Multi-omic characterisation of Barrett's oesophagus reveals a molecular continuum in the progression to oesophageal adenocarcinoma

Annalise Catherine Katz-Summercorn

Fitzwilliam College, University of Cambridge

January 2020

This dissertation is submitted for the degree of Doctor of Philosophy.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the Doctor of Philosophy Degree Committee.

Annalise C. Katz-Summercorn

January 2020

Summary

Multi-omic characterisation of Barrett's oesophagus reveals a molecular continuum in the progression to oesophageal adenocarcinoma

Annalise Catherine Katz-Summercorn

Barrett's oesophagus (BE) is the main risk factor for the development of oesophageal adenocarcinoma (OAC) yet only 0.4%/year of non-dysplastic (ND) BE cases progress to cancer (Bhat et al., 2011).

In this thesis I present the first comprehensive, multi-omics study comparing indolent, non-progressors with those who progressed to a range of dysplasia grades.

BE is highly heterogeneous with regards to mutational load, copy-number aberrations (CNAs) and structural variants (SVs). Mutational signatures are laid down early and persist regardless of progression status. Hence, Cosmic signature 17 (T:A>G:C in a CTT context), the hallmark of OAC, is visible in indolent, ND samples. *TP53* mutation, *GATA6* amplification, *ERBB2* amplification, *APC* mutation and whole genome doubling are confined to cases that have progressed with *TP53* being by far the most prevalent. In contrast *CDKN2A* alteration occurs early in around 50% of indolent cases. SV analysis reveals a dominance of translocations from the early ND grade.

Spatial analysis of multiple samples from across BE segments shows that the total mutation burden, total CNAs and total numbers of SVs to be surprisingly constant. However, clonality analysis highlights the complexity of BE, with some cases arising from a clear ancestral clone while others having minimal sharing of variants and showing heterogeneity in terms of driver events.

Transcriptomic analysis reveals a clear but gradual differential gene expression between ND and dysplastic biopsies. We demonstrate a loss of the intestinal metaplasia phenotype with progression and downregulation of the *HNF4A* pathway, a transcription factor with roles in intestinal development. In progression there is an upregulation of genes in the *ERK/MAPK* pathway.

In summary, BE is a heterogeneous disease that shows a continuum of abnormalities in the progression towards cancer. We hypothesise that it is the accumulation of events that tip the

balance to progression, rather than a stepwise model through the phenotypic dysplasia grades. Selected features could help to differentiate indolent from high risk disease and further work is being performed to identify scoring systems and biomarkers to apply in the clinical setting.

Contents

List of figures

List of tables

Collaborations

Ginny Devonshire and Lawrence Bower, Fitzgerald lab, performed all the initial processing of sequencing data: QC, alignment and running the bioinformatics pipelines for variant calling.

Dr Sriganesh Jammula, a computational post-doc in the Fitzgerald lab, guided me through the bioinformatics analyses. He, specifically, ran the software and did the annotations for the mutational and SV signatures, whole genome duplication and timing, chromothripsis, chromosomal instability, immune signatures and the point mutation driver gene de novo analyses. He created the circos plots and the copy number clustering and driver gene plots. He also ran the decision tree package.

David Wedge's lab, Big Data Institute, Oxford, were external collaborators working with us to call clonal and subclonal copy number aberrations in the cohort with Battenberg. Battenberg and DPClust were run by post-docs Anna Frangou and Iliana Peneva and they constructed the phylogenetic trees from this output for the heterogeneity analysis.

Acknowledgments

A huge thank you to:

Rebecca, for your guidance and support throughout my PhD.

Ganesh, it has been a pleasure working with such a phenomenal Bioinformatician, thank you for teaching me Bioinformatics!

CRUK for funding me.

Tissue Bank, for cutting many, many H&Es and kindly letting me add extra cases midweek.

The British Research Council and the Cambridge Experimental Cancer Medicine Centre for Barrett's tissue collection.

Maria, Monika and Shalini, for the endless boxes of frozen biopsies I asked you to grade.

GI Nurses, Tara, Bincy, Irene, Nicola, Rachel, and endoscopists, Massi, Wlad, for collecting all those samples and helping me to locate missing clinical data.

Ginny, you always remain unflummoxed despite how many different IDs I gave my samples and how many files I asked you to transfer!

Karol, for sitting next to me for 3 years, troubleshooting and teaching me how to think like a scientist. And how to use Google…

John, for being such a good and patient teacher when showing me the multiplex IHC.

Xiao, for your ideas and guidance.

Paul, for all your ideas on directions for this project.

Alex, Jason and Nuria for logging and storing so many biopsies and taking them to Tissue Bank. And Alex for taking pity on me in the -80 freezers and coming to help.

Babs, for helping me navigate OCCAMS data.

The Oxford team: David, Anna and Iliana, for the clonality analysis and your ideas for project directions.

Calvin, for your incredible skills at over-hauling the databases in the recent months. I wish that you had been here from the start!

Sujath and Adrienn, for your help with the RNA library preparation.

Shona, for always keeping cool and calm despite what we all went to you with.

Amber, for keeping me level headed!

Everyone else in the lab and all the staff and other groups at the Hutch. You've made it a wonderful place to work.

Staff at Starbucks – I won't name you all though I think that I could. I shall not estimate how much my PhD cost in Flat Whites.

Finally, to my husband, Dan, for all your support and encouragement. And to my son, Douglas, for sleeping well.

I could not have done this without you all.

Abbreviations

AF	Allele frequency
APC	Argon Plasma Coagulation
BE	Barrett's oEsophagus
BMI	Body mass index
CCF	Cancer cell fraction
CIN	Chromosomal instability
CN	Copy number
CNA	Copy number aberration
CT	Computer tomography
D2	Duodenum (second part)
DDR	DNA damage repair
EMR	Endoscopic mucosal resection
ESD	Endoscopic submucosal dissection
FC	Fold change
FISH	Fluorescence *in situ* hybridisation
GC	Gastric cardia
GM	Gastric metaplasia
GORD	Gastro-oesophageal reflux disease
GSVA	Gene set variation analysis
HGD	High grade dysplasia
IHC	Immunohistochemistry
IM	Intestinal metaplasia
IMC	Intramucosal carcinoma
ITH	Intra-tumoural heterogeneity
LGD	Low grade dysplasia
LN	Lymph node
LOH	Loss of heterozygosity
ND	Non-dysplastic
ND-NP	Non-dysplastic, non-progressor
ND-PP	Non-dysplastic, pre-progressor
NDBE	Non-dysplastic Barrett's oEsophagus
NE	Normal oEsophagus
NSAID	Non-steroidal anti-inflammatory
OAC	Oesophageal adenocarcinoma
PCA	Principle component analysis
PDT	Photodynamic therapy
PET	Positron emission tomography
RFA	Radiofrequency ablation
RIN	RNA integrity number
SBS	Single base substitution
SCC	Squamous cell carcinoma
SNV	Single nucleotide variant
SV	Structural variant
TMB	Total mutation burden
WES	Whole exome sequencing
WGD	Whole genome duplication
WGS	Whole genome sequencing
Wt	Wild type

1. Introduction

1.1 Oesophageal adenocarcinoma epidemiology, presentation and management

Oesophageal adenocarcinoma (OAC) is a malignancy arising from glandular epithelium usually occurring in the lower third of the oesophagus in proximity to the gastro-oesophageal junction. Although worldwide it is less common than its counterpart, squamous cell carcinoma (SCC), over the last 4 decades its incidence has been climbing, with more than a 6-fold increase, and it has surpassed the incidence of SCC in a number of Western countries (Coleman et al., 2018). In the UK, approximately 9000 cases of oesophageal cancer are diagnosed each year, of which 55% are adenocarcinomas (Cancer Research UK). It is thought to develop from Barrett's oesophagus (BE), a well-described, pre-malignant lesion which extends proximally from the gastro-oesophageal junction.

Clinically, OAC most commonly presents with a progressive difficulty in swallowing, first to solids and later to liquids, persistent heartburn, weight loss, anorexia and fatigue. In 2016, of all OAC diagnoses in the UK, 65.2% were via referral from a GP but 13.7% presented via an emergency admission to hospital (Varagunam et al., 2017). Presenting via an emergency admission has a very poor prognosis. However, the outcome is not much better in the former scenario because once the tumour is symptomatic there is a high probability of spread to lymph nodes. Overall, approximately 60% of patients have advanced, palliative disease at presentation. (Varagunam et al., 2017). This failure to diagnose oesophageal cancer in its earlier stages results in an overall 5-year survival of only 20% in Western countries (Coleman et al., 2018). In contrast, only 0.5% of diagnoses are made during endoscopic surveillance for BE (Varagunam et al., 2017). Meaning that not only is it generally diagnosed at an advanced stage, but also that we are not very good at finding the 'at risk' population.

OAC is diagnosed by the endoscopic appearance of the oesophagus coupled with histopathological analysis of biopsies. Further radiological investigations including CT and PET are then performed to determine the local invasion (T stage), lymph node involvement (N stage) and distant spread (M stage) of the disease (Figure 1, taken from (Thrumurthy et al., 2019). These are combined to give an overall staging of the tumour according to the Union for International Cancer Control-American Joint Committee on Cancer (UICC-

AJCC) TNM staging 8th edition (Amin et al., 2017) (Table 1). This staging is a guide to the likely 5-year survival. If the cancer can be diagnosed early at stage T1 or Tis (in situ carcinoma), the 5-year survival is >80% (Coleman et al., 2018).

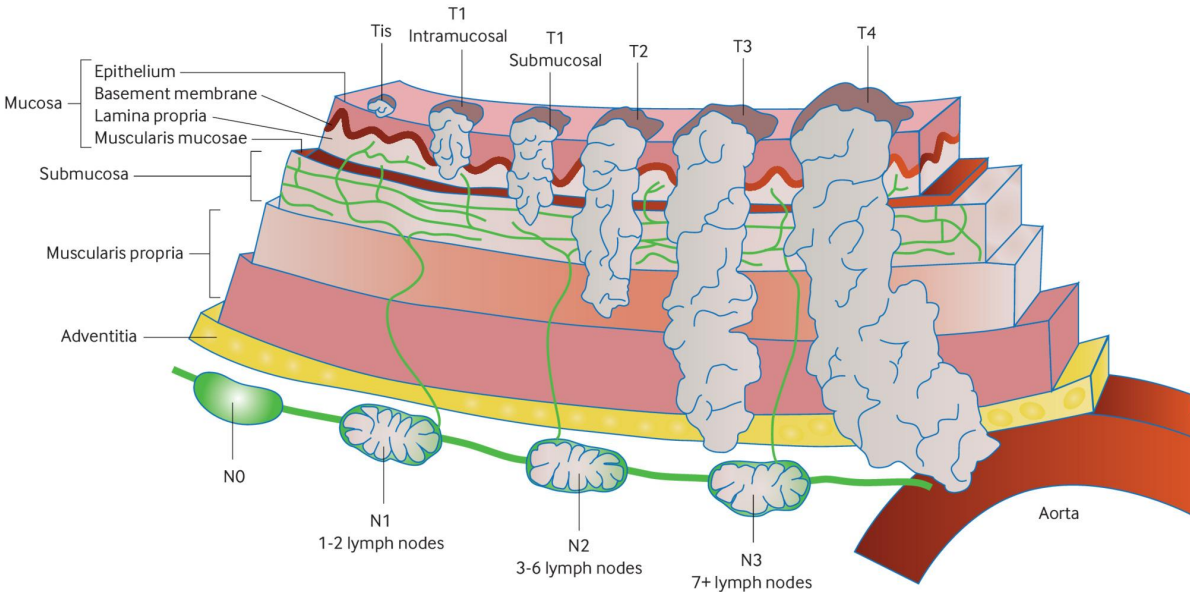


Figure 1 Definition of TNM staging of oesophageal adenocarcinoma

T (tumour) stages is-4 refer to the depth of invasion of the tumour through the wall of the oesophagus. Tis (in situ) has not invaded through the basement membrane of the epithelium. T1a (intramucosal) is confined to the mucosa, and T1b (submucosal) to the submucosa. T2: invasion through the submucosa; T3: full thickness invasion; T4: invasion into adjacent structures. N (nodal) stage is defined by the number of local lymph nodes involved. Referenced from Thrumurthy et al., BMJ 2019.

Clinical stage	T	N	M	Management	5 year-survival
0	Tis	N0	M0	Endoscopic	>90%
I	T1a, T1b	N0	M0	Endoscopic/ surgical	80.5%
II A	T1	N1	M0	Surgical with neoadjuvant CRT	45.1%
I B	T2	N0	M0	Surgical with neoadjuvant CRT	-
III	T2	N1	M0	Surgical with neoadjuvant CRT	17.6%
	T3, T4a	N0, N1	M0	Surgical with neoadjuvant CRT	-
IV A	T1-T4a	N2	M0	Surgical with neoadjuvant CRT	2.1%
	T4b	N0-2	M0	Palliative chemotherapy	
	Any T	N3	M0	Palliative chemotherapy	
IV B	Any T	Any N	M1	Palliative chemotherapy	-

Table 1 TNM staging, management and survival of oesophageal adenocarcinoma

Clinical staging of oesophageal adenocarcinoma according to the pathological extent of the tumour through the wall of the oesophagus (T), the involvement of lymph nodes (N) and the presence of distant metastases (M). CRT = chemoradiotherapy. 5-year survival per stage taken from Coleman et al., *Gastro*, 2018.

The tumour stage will also determine whether the disease can be managed with curative intent or not. Endoscopic therapy has advanced significantly in the past 10 years, where previously oesophagectomy was the only effective treatment for early neoplastic lesions. Early tumours confined to the mucosa can be resected endoscopically by endoscopic resection of the mucosa (EMR) or submucosa (ESD) as the risk of lymph node (LN) involvement is small. The risk of LN involvement for T1b tumours has been shown to depend on the depth of the invasion into the submucosa (sm1 = limited to the superficial submucosa 6%; sm3 = invading >500um 58%) (Gockel et al., 2011). ESD +/- adjuvant therapy is now recommended for early T1b tumours (Othman et al., 2019) as ESD was shown to offer the same 5 year survival as surgery (Gong et al., 2017). ESD of sm2 and 3 tumours are generally reserved for those unable to cope with the morbidity associated with an oesophagectomy. More advanced tumours with local nodal involvement are resected surgically with the local LN stations with neoadjuvant chemoradiotherapy. However, if the tumour has invaded adjacent structures, or in the presence of distant nodal disease or metastases, surgical resection is not possible and palliative chemotherapy is used. But 5-year survival for this stage IV disease is low at 2.1% (Coleman et al., 2018). If OAC could routinely be detected at Stages 0 or 1, survival would improve dramatically to >80%.

1.2 Barrett's oesophagus epidemiology, surveillance and management

BE is considered to be the main risk factor for the development of OAC and is thought to have a prevalence of 1-2% in the general population (Ronkainen et al., 2005; Zagari et al., 2008). However, it has been estimated from screening studies that most cases go undiagnosed. Studies have found 10-15% of people with GORD are found to have BE on endoscopy, although risk is affected by the severity and duration of the symptoms (Lieberman et al., 1997; Westhoff et al., 2005). The main other risk factors are being male, obesity, age >50 and white race (Edelstein et al., 2009). Currently, the British Society of Gastroenterology does not recommend unselected population screening, but endoscopy can be considered in patients with chronic reflux and the above risk factors (Fitzgerald et al., 2014).

BE is a metaplastic condition in which there is a change from the normal squamous epithelial lining of the oesophagus to a columnar epithelium. It is thought to develop in a step-wise process, with increased risk of further progression at each stage: from a non-dysplastic (ND) glandular epithelium, to low grade dysplasia (LGD; risk of progression 11.6%/year (Phoa et al., 2014)), high grade dysplasia (HGD; 19%/year (Wani et al., 2009)), and then intramucosal carcinoma (IMC; adenocarcinoma confined to the mucosa) (Figure 2).

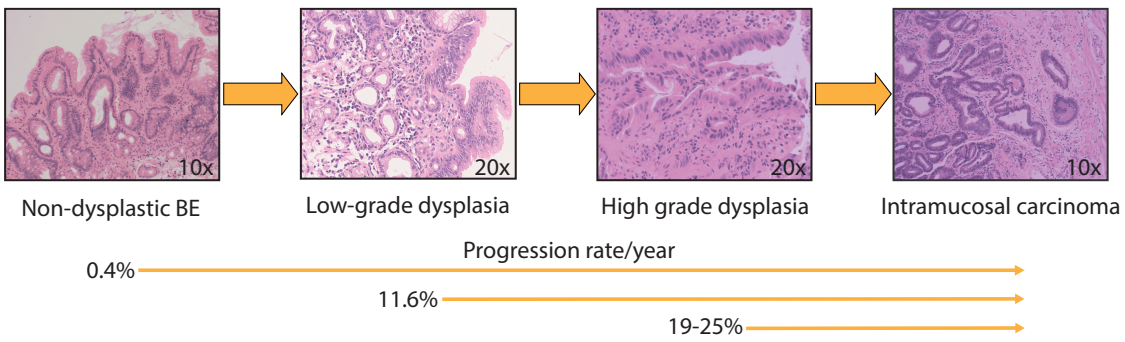


Figure 2 Grades of Barrett's oesophagus and rates of progression

Haematoxylin and eosin-stained sections. BE = Barrett's oesophagus. Rates of progression taken from Bhat et al. 2011, Phoa et al., 2014 and Shaheen et al., 2009. Image magnifications detailed.

However, only 0.4%/year of NDBE will progress (Bhat et al., 2011; Desai et al., 2012) and it is not understood what drives this progression. Therefore, currently, all patients with BE are monitored endoscopically every 2-5 years. The Seattle Protocol recommends that quadrantic biopsies be taken every 2cm along the length of the segment to look for the presence of dysplasia or IMC. The need for such regular sampling of the BE segment is because dysplastic lesions are often flat and cannot be distinguished from the surrounding BE under white light endoscopy. Although the biopsies are taken systematically, there is the potential to miss focal areas of disease (sampling bias). Biopsies are assessed histopathologically for the presence of dysplasia. p53 immunostaining may be used as an adjunct to aid diagnosis and now features in the British and European guidelines (Bas Weusten et al., 2017; Fitzgerald et al., 2014) but is not yet recommended in the US (Shaheen et al., 2016). The evidence for dysplasia and p53 as biomarkers will be discussed further later in the chapter.

Lesions with HGD or IMC are managed endoscopically. They can be resected by EMR or ESD and/or the area ablated by one of several methods including radiofrequency ablation (RFA), argon plasma coagulation (APC), photodynamic therapy (PDT) and cryotherapy. Randomised clinical trial data shows that treatment significantly reduces the cancer risk in patients with BE dysplasia: 2.4% vs 19.0% progression from HGD in the RFA group vs sham (at 12 months) (P=0.04) (Shaheen et al., 2009) and 1.5% vs 8.8% progression from LGD in RFA group vs surveillance (median follow-up 36 months) (Phoa et al., 2014). These studies have led to a change in practice so that intervention is routine for confirmed LGD (Fitzgerald, 2015; Fitzgerald et al., 2014; NICE, 2014).

1.3 The genomics of oesophageal adenocarcinoma and Barrett's oesophagus

1.3.1 Oesophageal adenocarcinoma is highly mutated, rearranged and driven by copy number aberrations

In order to understand the progression of BE, it is important to first consider the end stage of the progression, OAC. Over the last 10 years, with the technological advances of genomic sequencing, there have been huge international efforts to understand the molecular changes which confer a proliferative advantage to a cell and, therefore, drive tumourigenesis. The International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) spearheaded this with the genomic and exomic sequencing comparisons of large numbers of tumours. This led to many insights into the disease, showing that in OAC, somatic point mutations seem to be only a small part of the full picture and copy number aberrations (CNAs) and structural variants (SVs) are key features of the disease. However, this sequencing, whilst identifying new low frequency alterations in new oncogenic drivers, it has not identified any occurring at high frequencies. Instead it has highlighted the inter and intra-tumoural variation of OAC.

OAC is a highly mutated, heterogeneous disease, with a median of 6.4 mutations/Mb (single nucleotide variants (SNVs) and indels) (Frankell et al., 2019). This is possibly due to the mutagenic environment of acid and bile that the lower oesophagus is subjected to in gastro-oesophageal reflux disease. However, the majority of these mutations are synonymous, with a mean rate of only 151.4 non-synonymous somatic variants per genome. Disappointingly, the only gene found in these studies to be highly recurrently affected by point mutation is *TP53* in 72% of cases (Frankell et al., 2019); already known to be the case from previous candidate gene studies from more than 20 years ago (Moore et al., 1994; Neshat et al., 1994; Prevo et al., 1999; Schneider et al., 1996). OAC is also dominated by aneuploidy and CNAs, which has long been known from single nucleotide polymorphism (SNP) array studies (Bandla et al., 2012; Frankel et al., 2014; Paulson et al., 2009) and more recently from WGS data e.g. (Secrier et al., 2016). Integration of these analyses with expression data has facilitated the identification of genes likely to be drivers due to alterations in expression by amplification or deletion. For example, when both point mutation and deletion were considered, *CDKN2A* was altered in 28% of tumours. Known oncogenes *KRAS, MYC* and

ERBB2 were amplified each in approximately 20% of cases (Frankell et al., 2019). Overall, 76 driver genes have been discovered in OAC, with a median of five events in driver genes per cancer. Only 1% of cancers had no driver events identified but it is likely that we are still missing some of the lowest frequency events, especially in genes driven by CNA.

Whole genome duplication (WGD), the acquisition of a complete second set of chromosomes within a cell, is a common phenomenon in cancer, thought to be due to errors in a number of mechanisms during mitosis e.g. cytokinesis failure. It has been shown to be more frequent in tumours with higher proliferation rates and has been associated with *TP53* mutation: a normally-functioning p53 prevents a WGD cell from entering the cell cycle and proliferating (Bielski et al., 2018). Chronologically, *TP53* mutation occurs prior to WGD and this has been shown to be the case in 97.3% of patients across cancer types (Bielski et al., 2018). WGD has been observed in 50-62.5% of OAC (Secrier et al., 2016; Stachler et al., 2015).

OAC has also been shown to be dominated by large scale structural rearrangements (structural variants, SVs) with a median of 263 SVs per tumour (Nones et al., 2014; Secrier et al., 2016; TCGA et al., 2017). These large-scale rearrangements can accumulate slowly over time, due to mitotic segregation errors, or rapidly by catastrophic evolutionary events of 'chromothripsis' (chromosome shattering). Chromothripsis has been exhibited in 30% of tumours (Nones et al., 2014).

In addition to these signatures, 31% of OAC tumours exhibit a phenomenon called 'kataegis', characterised by localised hypermutation dominated by C>T and C>G mutations. The exact importance of these events is unclear (Nones et al., 2014).

1.3.2 Mutational Signatures in oesophageal adenocarcinoma

Somatic mutations seen in individual tumours can be categorised into different combinations to further understand the biological processes and the exogenous and endogenous exposures generating the mutations. Rather than observing the specific genes which are altered by the mutations, the genome-wide shifts in base substitutions are analysed. There are 6 classes of single base substitution (SBS): C>A, C>G, C>T, T>A, T>C, T>G and these are considered in the context of their flanking 3' and 5' bases within the trinucleotide. This results in 96 permutations of which the proportions of each within the genome are classified into signatures. 20 distinct signatures were initially described (Alexandrov et al., 2013) and this

has recently increased to 67 (https://cancer.sanger.ac.uk/cosmic/signatures/SBS/). In an analysis of WGS data on 129 chemo-naïve cases of OAC, six signatures were prominent (Figure 3).

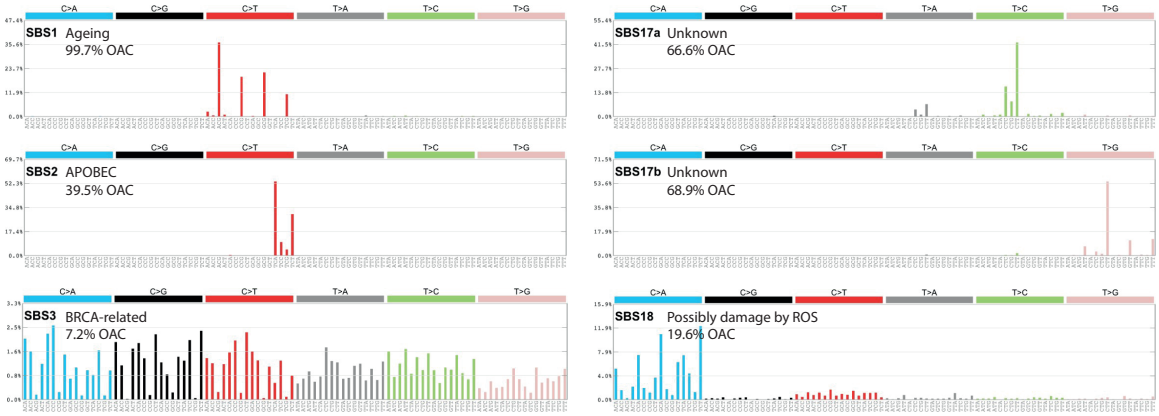


Figure 3 Six mutational signatures identified in oesophageal adenocarcinoma

Single base substitution (SBS) signatures described by Alexandrov et al., 2013 and shown to be prominent in oesophageal adenocarcinoma (OAC) (Secrier et al., 2016). ROS = reactive oxygen species. Percentages in OAC taken from TCGA data via COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures/).

SBS17 is dominated by T>G substitutions in a CTT context. It was originally thought to be linked with acid reflux, however potential new causative factors are now being considered including reactive oxygen species (Pich et al., 2019). It is also seen following exposure to exogenous 5-fluorouracil (Christensen et al., 2019). It is considered the hallmark signature in OAC and has a variant of the signature with a higher frequency of T>C substitutions termed SBS17B. SBS3 is caused by defects in the BRCA homologous recombination pathway resulting in a complex set of mutations. SBS1 is the ageing signature, characterised by C>G mutations in a *CG context. SBS2 is dominated by C>T mutations in a TCA/TCT context from APOBEC-driven hypermutation. Finally, SBS18-like signature, previously described in breast and stomach cancers is a C>A/T substitution in a GCA/TCT context. In the analysis they considered the dominant contribution of these 6 signatures per case and identified three subgroups which suggested differing aetiology and differences in genomic stability. The three groups are: the DNA damage repair (DDR) impaired group (18%) which is dominated by the BRCA signature and has a high degree of genomic instability; the mutagenic subgroup (53%) which is dominated by S17 and has a significantly higher mutation rate than the other groups; and the C>A/T dominant subgroup (29%) which is predominantly the age and SBS18-like signatures and has the lowest rate of genomic

instability (Secrier et al., 2016). These subgroups have led to new hypotheses for therapy: e.g. mutagenic tumours, with a higher mutational burden and neoantigen load, may be more responsive to immunotherapy; and DNA damage repair (DDR)-impaired tumours may respond to treatment with PARP (poly ADP ribose polymerase) inhibitors.

In summary, whilst OAC is highly mutated and heterogenous between patients, it seems to be dominated by copy number alterations in genes which drive progression of the disease. Large cohorts have been needed to identify these altered genes because of the low frequencies of recurrent alterations. Whole genome doubling and larger structural rearrangements are also seen. Algorithms can be used to predict the timing of events, but we do not know when these key alterations occur in the evolution of the disease. Studying Barrett's oesophagus allows the possibility to investigate when these alterations occur pre-progression.

1.3.3 Barrett's oesophagus is highly mutated and affected by copy number changes

The molecular changes which drive the progression of non-dysplastic Barrett's oesophagus (NDBE) to cancer remain incompletely understood despite the last 30 years of research.

The genomic instability of BE was first demonstrated in the late 1980s using flow cytometry. Aneuploidy or an increased 4N fraction were associated with dysplastic and adenocarcinoma biopsies (Reid et al., 1987). These findings at baseline endoscopy also predicted progression (Reid et al., 1992, 2000). As technology improved, the importance of the accumulation of focal chromosomal aberrations during the metaplasia-dysplasia-carcinoma sequence was further shown using comparative genomic hybridisation, candidate region analysis and low-density SNP arrays (Paulson et al., 2009; Riegman et al., 2001). But apart from the consistent early deletion of 9p21 and the late loss of heterozygosity of *TP53*, other results were highly heterogeneous, as seen in the cancer.

Later, whole-genome, high-density SNP array studies were able to focus in on these changes at a higher resolution. One study considered the different stages of progression showing that in ND, LGD, HGD and OAC, the mean percentages of SNPs with allele loss increased: 0.1%, 1.8%, 6.6% and 17.2% respectively (Gu et al., 2010). There were many recurrent small regions of loss in the later stages, disrupting single genes, predominantly at fragile sites, in addition to the expected losses of the loci containing *TP53* and *CDKN2A*. 17p (containing

TP53) was lost in 5.3% of LGD, 25.0% HGD and 47.6% of OAC. 9p21 (the locus for *CDKN2A*) was lost in 68.4% of LGD, but only 5% of ND. However, both earlier and more recent studies have shown *CDKN2A* to be more frequently lost or mutated at the ND stage and not predictive of progression to OAC (Galipeau et al., 2007; Reid et al., 2001; Weaver et al., 2014).

SNP arrays have also been used to compare the accumulation of CNAs over time in patients who go on to progress to cancer, versus those that do not. Genomes of non-progressors show very little genomic diversity and stable genomes over time. Conversely, in the progressors the genomes evolve significantly as they approach the cancer time point, with increased copy number variation, predominantly in the preceding 48 months prior to cancer diagnosis (Li et al., 2014).

Most recently, the focus has been on WGS and whole exome sequencing studies (WES). Whilst efforts have mostly concentrated on the cancer, two studies sequenced the BE lying adjacent to OAC. They found this BE to also be highly mutated and heterogeneous, like OAC. The mutation rates in these cases of BE adjacent to cancer were, surprisingly, higher than that of many other invasive cancers: 1.3-6.76 mutations/Mb (Ross-Innes et al., 2015a; Stachler et al., 2015). The cohorts in these studies were small (n = 23-25) and the pathology of the BE samples predominantly ND. Of these, dysplastic BE appeared to have a higher mutation rate than the ND on WES, but this was not demonstrable at lower depth 50X WGS. In contrast, the ND cases had very few copy number changes compared to the cancer, with predominantly diploid genomes (median % of the genome with CN 2 = 99.7%; cancer: 37.6%). Both studies found BE to be polyclonal, with surprisingly little overlap between the OAC and the adjacent BE (13/23 samples had <20% overlap of SNVs in one study). By sequencing multiple samples from individual patients, Ross-Innes et al were able to identify 6 distinct clones present in one patient. On considering the BE clonally related to the cancer, Stachler et al. demonstrated that *TP53* inactivation appeared to be an early event, followed by whole genome duplication (WGD) and subsequent further genomic instability. In their analysis, 62.5% of cases demonstrated this phenomenon. This concept of genome doubling could explain why some patients with BE progress rapidly to cancer.

Despite this above work, important questions remain regarding the BE adjacent to cancer. For example, what is the local effect of the cancer on the surrounding area and whether this adjacent BE is representative of the pre-progression stages. Furthermore, given that the majority of the clones seen in the BE are unrelated to the adjacent tumour, this may be BE

that never had the capacity to progress. In order to answer these questions, the pre-malignant stages of the disease need to be considered. An earlier study used a next-generation sequencing 26-gene panel, based on the mutations observed in OAC WGS data. They compared the frequency of SNVs in specific genes in BE from patients who had never progressed to dysplasia (median follow-up >8 years) with biopsies from patients with HGD (Weaver et al., 2014). In the ND cohort (n=40), 53% had a mutation within the biopsy, and 91% of the HGD cases (n=43) were mutated. It was surprising that a number of genes observed in HGD cases were mutated at low frequency in never-dysplastic BE, including *ARID1A*, *SMARC4A* and *CDKN2A*. Only *TP53* mutation distinguished between the two stages (2.5% NDBE, 72% HGD), highlighting its potential use as a stage-specific biomarker, confirming previous studies.

Since then, larger genomic panels have been used to compare the ND samples from patients who went on to progress and those that did not, introducing some uncertainty as to when *TP53* mutations occur (Del Portillo et al., 2015; Stachler et al., 2015). A study using a 243-gene panel found *TP53* mutations in the biopsies at a time-point prior to progression to HGD/OAC (46%), but only in 5% of the non-progressors. However, the non-progressor group included patients with LGD and many cases were re-reviewed and downgraded to NDBE, which raises the question as to whether cases with *TP53* mutations were downgraded appropriately. There was also no germ line comparison for mutation calling and *TP53* mutations were further manually curated. They did not observe significant differences in copy number profile or ploidy between the pre-progressors and the non-progressors ND samples either from the ploidy or in the more detailed analysis of CNAs. This finding was out of keeping with prior literature (Li et al., 2014).

Overall, there have not been any large WGS studies of BE cohorts considering the stages of progression. One of the problems with panels is that the choices of genes on them are based our knowledge of cancer and de novo discoveries cannot be performed. They also do not permit the analysis of structural rearrangements and whole genome CNAs. Without being able to look at these we are missing key features for understanding how these events interplay in the progression of this heterogenous disease.

1.3.4 Clonal diversity in Barrett's oesophagus

Another important consideration in the evolution of BE, that is becoming increasingly recognised, is how the clonal diversity of a segment relates to the risk of progression to

cancer. This was first shown in BE more than 10 years ago, when the Reid group adapted measures from ecology and evolution to quantify clonal diversity with the Shannon diversity index. The hypothesis was that the larger the number, and the more genetically different the clones within the BE segment were, the higher the potential for that segment to progress. Multiple biopsies per patient were assessed by flow cytometry, fluorescent in situ hybridisation (FISH) and *TP53* and *CDKN2A* sequencing. The number of clones, Shannon index and genetic divergence, based on loss of heterozygosity (LOH), were strongly predictive of increased progression to OAC (Maley et al., 2006). The number of clones alone was a slightly better predictor of progression than the Shannon diversity and, given the ease of measuring the number of clones in a neoplasm, it has been suggested as a more useful measure. Interestingly, there were no significant differences in the Shannon diversity between tissues with and without 17q (*TP53*) or 9q (*CDKN2A/p16*) LOH, nor was it proportional to time since loss of *TP53*.

FISH can also discriminate genetic diversity at a single-cell resolution and has been used on brush cytology specimens from BE patients. In a study comparing the baseline samples from progressors and non-progressors, genetic diversity could be seen on a single-cell basis, by scoring 50 cells based on four FISH probes. Across patients, loss of one *p16* allele was observed in 51% (163) of patients, with complete loss in 5%. Overall, the level of clonal diversity at baseline in ND was indicative of progression risk and, importantly, did not change significantly over time (Martinez et al., 2016). The size of the biggest clone was not a prognostic marker: further supporting the findings that it is the number of clones, rather than clone size that is predictive of progression. Furthermore, *p16* diversity was not a useful predictor of progression.

The diversity seen at an individual crypt level in BE, using SNP arrays to measure CNAs, correlates with that seen at the biopsy level and does not provide additional information about genetic diversity (Martinez et al., 2018). However, proximity to the GOJ correlates with increased genetic diversity – which could explain the high incidence of tumours developing in the BE at the GOJ, rather than proximally. The idea that the progression risk of BE can already be determined at baseline has the potential to massively reduce the need for BE surveillance if it can be successfully translated into a diagnostic test.

1.4 Expression analyses in Barrett's oesophagus

The focus of most RNA expression studies has been either looking for differentially expressed genes between OAC and NDBE or comparing NDBE to normal squamous oesophagus (NE). The main questions considered have been either about how the cancer develops or how BE develops in the squamous oesophagus.

To date there has been one whole transcriptome study of BE and dysplasia with small numbers: 17 NE, 14 NDBE, 8 LGD and 12 OAC. The focus of the paper was on the expression differences between OAC and NDBE, however they did demonstrate the upregulation of 6 transcription factors, in LGD compared to NDBE, involved in cell proliferation, differentiation and transformation processes (Maag et al., 2017). Specifically, *FOSB, NR4A1, EGR1, FOS, EGR3* and *ATF3*. An earlier study used gene expression from microarrays to find a 90-gene signature pattern which could differentiate HGD from NDBE (Varghese et al., 2015).

However, whilst there remains controversy over the cell of origin of BE (Que et al., 2019) there is little evidence to support a transdifferentiation of squamous epithelium. With stronger evidence for an origin from either the submucosal glands or gastric cells (Nowicki-Osuch et al., 2019 (unpublished); Owen et al., 2018). Therefore, using NE as the tissue for comparison, as the above study did, is probably misleading. Phenotypically, BE with intestinal metaplasia shares many features with duodenum, and also the pyloric glands of the stomach. The epithelium is a combination of foveolar epithelium and goblet cells. In BE with IM, the brush border containing enterocytes is absent and the BE gland features a mucinous base compartment, which produces bicarbonate, and a specific repertoire of mucins. So, duodenum and gastric are better tissues for comparison when trying to consider the altered expression in BE.

Expression data is key both for understanding the biological effects of genomic alterations but also for defining potential protein biomarkers from genes with increased expression. Whole transcriptome sequencing offers the further possibility of considering the non-coding RNAs in understanding the biology of progression. Overall, larger studies across the grades are needed.

Protein studies can directly consider the changes occurring with progression. Prior to expression analyses, these mainly focussed on specific pathways in cell lines. Proteomics has been used to give an overall, unbiased approach in cell lines (Breton et al., 2008) and

human tissue but with a focus on the differences between the BE and the OAC, rather than the stages of progression (Elsner et al., 2012; O'Neill et al., 2017; Peng et al., 2008; Streitz et al., 2005; Zhao et al., 2007). The low throughput of proteomics has, so far, limited its use for detecting upregulated proteins in such a heterogenous disease.

1.5 Epigenetic alterations in Barrett's oesophagus

Modification of gene expression by aberrant DNA methylation plays a fundamental role in cancer development. It is another mechanism by which the expression of tumour suppressor genes and oncogenes is altered and compliments the effects of SNVs and CNAs seen at a genomic level. Methylation of cytosine residues, by DNA methyltransferases, occurs at CpG sites: where the cytosine is linked to a guanine residue by a phosphate. Areas of the genome with a high density of CpG sites are termed CpG islands and the hypermethylation of these areas and promotor regions results in transcriptional silencing and decreased expression of genes. Conversely, hypomethylation causes overexpression.

Methylation-induced inactivation of *CDKN2A* is one of the commonest changes observed in BE metaplasia and it was the first gene found to be affected by methylation in BE in early gene-specific work (Wong et al., 1997). Methylation of its promoter region is seen in 15% of BE tissue, yet it is unmethylated in normal tissue (Hamilton et al., 2006). It seems to occur early in the process of progression and it is thought that the clonal expansion of p16-/- cells may form an environment conducive to the development of other genetic events, leading to OAC (Maley et al., 2004). There has been a lot of work to combine gene-specific methylation into panels in order to predict disease progression. For example, a 4-gene panel of *SLC22A18*, *PIGR*, *GJA12* and *RIN2* distinguished between NDBE and HGD/OAC with a 97% specificity and 94% sensitivity in a retrospective cohort (Alvi et al., 2013).

It has only been recently that the technology has developed to allow for an unbiased analysis across the whole epigenome, rather than focussing in on a limited number of CpG islands (27K arrays) and the differentially-methylated sites (Agarwal et al., 2012; Kaz et al., 2011; Xu et al., 2013). Using 450K arrays, OAC and BE have been shown to cluster together, with distinct separation from normal oesophagus (NE), suggesting that aberrant methylation is an early event in the stepwise progression of BE to OAC. However, as with the transcriptomics, this clustering could just represent the phenotypic differences between the glandular BE/OAC and squamous oesophagus. Stomach and duodenum are needed as comparisons to confirm if the methylation changes are really early events.

Hypermethylation is seen mainly within the CpG-rich promoters, with the regions outside being hypomethylated (shelf/gene-poor regions and the body of genes) (Krause et al., 2016). However, cohorts of BE to date have been small, with the focus predominantly on cancer. A study of 12 non-progressor BE and 12 progressor BE revealed global trends towards hypomethylation in the progressor group (Dilworth et al., 2019) although, oddly, they did not find a difference in CNAs between the two groups: a finding out of keeping with other studies. In a study of 125 OAC and 19 BE (11 taken at the cancer time-point), OAC/BE clustered into two distinct groups: a 'gastric-like' group and a 'CpG island methylator phenotype (CIMP)-like' group: a term first coined in colon adenocarcinoma. The methylation profile of the 'gastric-like' group was similar to that of gastric mucosa whereas the CIMP-like group were characterised by hypermethylation in the CpG islands. Interestingly, the top quantile of most hypermethylated tumours conferred a poor survival compared to all other tumours in the analysis. Methylation data was integrated with transcriptomics to show that 57% of testable sites correlated with gene expression changes (Krause et al., 2016). Epigenetic changes identified have the potential to be useful biomarkers in predicting disease stage and progression, especially if combined into integrated panels by using genomic features.

1.6 Integrated analyses

It is becoming clear from work in other cancers e.g. lung and oesophageal squamous cell carcinoma (Farshidfar et al., 2017; TCGA, 2012) that the real power of these sequencing methods comes when the analyses are integrated. For example, expression data can highlight the downstream effect of a mutation or copy number change, and methylation can explain expression changes where no mutation is seen. For OAC, the integration of WES with SNP-array profiling, DNA methylation profiling and mRNA/microRNA sequencing (TCGA et al., 2017) showed that CDKN2A, which is mutated in 15% of tumours (Secrier et al., 2016), was inactivated in 76% of OACs in total when mutation, deletion and epigenetic silencing were considered. And combining genomic with transcriptomic data has significantly increased the identification of driver events in OAC (Frankell et al., 2019). To date, however, analyses of this kind have mostly focussed on cancer, with only a few focussing on pre-malignant lesions: colorectal adenomas, hepatocellular adenomas and pre-invasive lung cancer (Druliner et al., 2018; Nault et al., 2017; Qu et al., 2016; Teixeira et al., 2019). This type of analysis could be particularly useful in a disease as heterogeneous as BE.

1.7 Clinical challenges and application to screening and surveillance

Ultimately, the reason the evolution of BE needs to be understood better is so that we can improve management and early detection of the disease. There are a number of clinical challenges in the current management of BE.

1.7.1 Diagnosing dysplasia

Firstly, the histopathological diagnosis of dysplasia is the current gold standard for risk of progression of BE to OAC, but it is not perfect. Despite significant, quality evidence to support its use (Phoa et al., 2014; Shaheen et al., 2009) it does not excel as a biomarker because its diagnosis is so subjective, with inter-observer variability, and there is a propensity for overdiagnosis. The Sharma group found a kappa coefficient of 0.11 (95% CI 0.004-0.15; none to slight agreement) between 3 pathologists for diagnosing LGD, clearly showing just how difficult it can be (Vennalaganti et al., 2017). A Dutch group demonstrated the likelihood of overdiagnosis by taking 293 LGD biopsies, diagnosed by a pathologist in the clinical setting, and subjecting them to review by an expert panel of pathologists. A high percentage, 73%, of the biopsies were downgraded to ND or indefinite for dysplasia. Of those which were confirmed to be LGD, there was a 9.1% per patient-year risk of progression. But this fell to only 0.6% and 0.9% per patient-year for the downgraded NDBE and indefinite biopsies respectively (Duits et al., 2015), clearly showing that dysplasia is an excellent predictor of risk, but only if correctly diagnosed.

The only clinically recognised adjunct to dysplasia diagnosis is p53 immunohistochemistry and it has been widely studied as a biomarker. Protein accumulation can occur when *TP53* mutation in one of the alleles results in an increased half-life of the protein by stabilizing it and preventing degradation. This accumulation of p53 has been shown to precede development of HGD/OAC by several years (Davelaar et al., 2015), an important characteristic for a potential biomarker. Sikkema et al. found p53 over-expression to result in a five-fold increased risk of progression to HGD or OAC, independent of the presence of LGD (95% CI 2-14.5, p=0.004) (Sikkema et al., 2009), and other studies have shown to predict progression from LGD to HGD/OAC, with a 63.6-100% sensitivity and 68-93% specificity (Davelaar et al., 2015; Kaye et al., 2009; Murray et al., 2006; Skacel et al., 2002; Weston et al., 2001; Younes et al., 1997). Since then, it has been realized that not all

mutations stabilize the protein: they may truncate it or result in non-expression, and so the absence of staining for p53 has been recognized to also have clinical utility. The key studies looking at p53 are summarised in Table 2 (Bird-Lieberman et al., 2012; Davelaar et al., 2015; Galipeau et al., 2007; Kastelein et al., 2013; Kaye et al., 2009; Reid et al., 2001; Sikkema et al., 2009; Skacel et al., 2002; Weston et al., 2001; Younes et al., 1997).

Biomarker	**Reference**	**EDRN stage**	**Sample size**	**Finding**
TP53 LOH using flow cytometry	Reid et al., 2001	Phase 3/4: prospectively collected samples, retrospective analysis	325	RR=16, p<0.001
	Galipeau et al., 2007	Phase 3/4: prospectively collected samples, retrospective analysis	243	RR=10.6 (95% CI 5.2-21.3, p<0.001)
P53 positive on IHC	Younes et al., 1997	Phase 3 retrospective	5 progressors, 25 non-progressors	Correlates with progression from LGD to HGD/OAC p=0.0108. 100% sensitivity, 93% specificity in predicting progression.
	Weston et al., 2001	Phase 4 prospective	5 progressors, 43 non-progressors	Kaplan-Meier curves differed significantly betweenp53 positive and negative patients with progression from LGD.
	Skacel et al., 2002	Phase 3 retrospective	8 progressors, 8 non-progressors	Correlates with progression from LGD to HGD/OAC p=0.017. 88% sensitivity, 75% specificity in predicting progression.
	Kaye et al., 2009	Phase 3 retrospective	154 progressors, 32 non-progressors	80% sensitivity, 68% specificity in predicting progression.
	Sikkema et al., 2009	Phase 4 prospective	27 progressors, 27 non-progressors	HR 6.5 (95% CI 2.5-17.1)
	Kastelein et al., 2013	Phase 3/4 prospectively collected samples, retrospective analysis	49 progressors, 586 non-progressors	P53 over-expression: RR 5.6, 95% CI 3.1 to 10.3. Loss of p53 expression: RR 14.0, 95% CI 5.3 to 37.2

	Bird-Lieberman et al., 2012	Phase 3/4 prospectively collected samples, retrospective analysis	Nested-case control. 89 progressors, 291 non-progressors	Risk of OAC alone OR 1.95, 95% CI 1.04-3.67. P53 was not found to predict HGD/OAC progression in multivariate analysis.
	Davelaar et al., 2015	Phase 4 prospective	116 patients	Progression to HGD/OAC 17 (95% CI 3.2-96, p=0.001. Progression to HGD only: OR 30.8 95% CI 3.78-308, p=0.002. IHC showed increased sensitivity (81.8%) but decreased specificity (85%) for progression to HGD when combined with FISH LOH.

Table 2 Studies of p53 as a biomarker in predicting progression

LOH = loss of heterozygosity, IHC = immunohistochemistry, FISH = fluorescent in-situ hybridisation, CI = confidence interval, RR = relative risk, HR = hazard ratio, EDRN = Early Detection Research Network, LGD = low grade dysplasia, HGD = high grade dysplasia, OAC = oesophageal carcinoma.

Other features of *TP53*, whilst not used clinically, also predict progression risk or likelihood of dysplasia. Loss of heterozygosity (LOH) for *TP53* was first shown in 2001 to be associated with a 16-fold increased risk of progression to OAC from NDBE (Reid et al., 2001). However, LOH detection requires multiple technical steps making it not easily applicable to routine clinical use. *TP53* mutation, as described earlier in the chapter has strong potential as a biomarker given that it is stage-specific: 2.5% never-dysplastic BE (n=66); 72% BE with HGD (n=43) ($p<0.0001$) (Weaver et al., 2014). However, on its own is not sufficiently sensitive given that we have shown that it is only mutated in 72% of dysplastic samples. Combination of *TP53* mutation into biomarker panels may overcome this. Overall, to date, many studies have attempted to combine biomarkers into panels for diagnosing dysplasia (Eluri et al., 2018; di Pietro et al., 2015) but many of them include diagnosis of dysplasia in their prediction of progression (Duits et al., 2019; Parasa et al., 2018). None so far have been promising enough to be further developed for clinical use.

1.7.2 Prediction of progression and endoscopy for surveillance

The other main challenges in management are around the need for long-term surveillance with endoscopy. We are not able to predict the 0.4%/year of patients with NDBE who will progress to OAC. This puts a huge burden on the NHS in following these patients for many years. Furthermore, with no routine screening of the population for BE, only 7% of patients with BE and OAC have had their BE diagnosed in advance (Bhat et al., 2015). In addition, endoscopy, whilst the only recommended method for surveillance, has a number of limitations. It is expensive, time-consuming and unpleasant for patients but, more importantly, biopsies can miss focal areas of dysplasia or cancer (sampling bias).

One way of overcoming sampling bias is moving towards cytological brush sampling methods e.g. the Cytosponge^{TM} which can be used in the community (Ross-Innes et al., 2015a). These sample cells from the whole lining of the oesophagus but the structure of tissue is, somewhat, lost making diagnosis difficult except by experts in Cytology. Immunohistochemistry (IHC) staining using Trefoil Factor 3 (TFF3) (Ross-Innes et al., 2015b) or MUC2 (Zhou et al., 2019a) can diagnose the presence of BE, but currently these patients then need an endoscopy to investigate for dysplasia and enter into surveillance. This would put strain on an already busy National Health Service. Ideally, another level of test is needed to diagnose the presence of dysplasia on cytology or also to risk stratify patients and determine who is likely to progress.

P53 IHC and *TP53* mutation have been tested on Cytosponge samples as part of a panel with glandular atypia, aurora kinase A IHC, age, BE length and waist-hip ratio. All the patients in the low risk group were ND. 87% of patients in the high risk group were dysplastic, but the moderate risk group formed a large, mixed group who would still require an endoscopy (Ross-Innes et al., 2017). This showed the potential of applying a panel to FFPE cytology specimens. A commercial driver gene mutation panel has also been applied to a small cohort of FFPE Cytosponge samples to diagnose dysplasia with a 71.4% sensitivity and 90.3% specificity (Katz-Summercorn et al., 2017). These studies either used a few known biomarkers or large non-specific gene panels, highlighting the clinical need for an improved understanding of the key events driving the progression of BE.

To address this overall clinical need, we need to start off by better characterising the stages of BE, in order to better understand the heterogeneity of the disease and what drives the progression of BE to cancer.

Hypothesis and Aims

It was on this background of evidence that I formed the hypothesis that performing an integrated analysis of multiple sequencing modalities on the stages of Barrett's oesophagus (BE) would overcome the heterogeneity of the disease, and lead to new insights about the key biological processes driving the progression of non-dysplastic BE to cancer. I undertook the creation of a cohort of patients in order to address the following aims:

1. Elucidate the key biological processes driving BE to progress to oesophageal adenocarcinoma (OAC) by performing an integrated analysis of genomic and transcriptomic sequencing of the individual grades of BE.
2. Consider the heterogeneity and clonal evolution of BE segments and how this may influence progression.
3. Identify how the biological findings may be integrated with clinical information in order to categorise patients into high or low risk of progression.

2. Methods

2.1 Cohort design

All Barrett's oesophagus (BE) patients were selected from our Biomarker and BEST2 Research Databases (Cell Determinants Biomarker Study: REC no. 01/149, BEST2 study: REC no. 10/H0308/71, Case1 study: Commercial; Rec. no. 14/EE/0015) for whom snap-frozen samples had been collected at endoscopy. This biopsy sampling at endoscopy included both the strategic sampling of the Seattle protocol (quadrantic biopsies every 2cm) and targeted biopsy of raised lesions/abnormal areas on narrow band and autofluorescence imaging. Biopsies were frozen in liquid nitrogen in the endoscopy room and then stored on dry ice for transfer. Histology reports of FFPE diagnostic biopsies were used to identify patients with the different grades of disease. Patients were excluded if they had progressed past the grade of interest, either before or at a later date. This was in order to be absolutely certain of the grade being sequenced and negate the risk of local effects from prior higher grades. Equally, samples with imminent future higher grade of progression were excluded because of the possibility that the higher grade was already present and missed at Barrett's surveillance. Patients were also excluded if they had received previous ablative treatment of their BE. Biopsies representing the independent grades could not be adjacent to cancer. Non-progressor patients with long follow-ups and long segments were selected where possible and pre-progressor samples were taken as far in advance of progression as available. In addition, cases of BE adjacent to cancer (Trio BE: cancer-BE-normal) were selected as a comparison (OCCAMS Rec. no. 10-H0305-1). These frozen samples were taken either from the oesophagectomy specimen or at the staging endoscopy. BE was sampled at the greatest distance possible from the tumour to avoid contamination (Table 3).

All biopsies underwent a strict, uniform pathology review process. A section from each frozen biopsy was cut and stained with haematoxylin and eosin (H&E) and reviewed by a consultant pathologist to assess the composition of the biopsy. For the main, pre-cancer BE cohort, any potentially suitable biopsies were then reviewed independently by a further 2 consultant pathologists. All pathologists were blinded to the grade of the patient. Sample grade was determined by an agreement of at least two pathologists. Samples with no agreement were reviewed by the 3 pathologists together to reach a consensus. Dysplastic

samples for sequencing had to have a pathological cellularity of dysplasia of >30% and were included even if the dysplasia was not all the highest grade the patient was known to have. Cellularity refers to the percentage of the tissue which is composed of dysplastic BE cells, with the rest composed of e.g. NDBE, stroma, squamous contamination and immune cells. A 30% minimal cellularity was used to try to ensure that 50X sequencing would cover the pathological mutations to an adequate depth but balanced against having enough samples in the cohort. Cellularity was assessed by the eye of the pathologist. Whilst it is known that this can lead to an overestimate in cellularity, the same method has previously proven successful in our lab for tumour sequencing (Secrier et al., 2016). NDBE biopsies had to contain intestinal metaplasia (IM). Samples with only gastric metaplasia were avoided because this phenotype of BE has an extremely low risk of progression and surveillance is not recommended for short-segment GM (Fitzgerald et al., 2014). For the Trio BE cohort, H&Es were reviewed independently by two pathologists and only dysplastic cases reviewed a third. Trio BE samples were excluded if there was any tumour contamination in the biopsy. Across the whole cohort, no squamous epithelium could be present in any sample however, samples with inflammation were not excluded. Duodenum was used as the germline reference as blood had generally not been collected. Where not available, blood was used if possible or normal squamous oesophagus (verified with H&E staining) in that order.

Essential inclusion criteria	**Preferable inclusion criteria**	**Exclusion criteria**
>= 30% cellularity for dysplasia in dysplastic samples, or IM in non-progressor samples	Long prior follow-up for non-progressors	Previous thermal ablative therapy
Consensus pathology review	Long overall follow-up	Prior higher grade
Snap frozen tissue		
Matched germline available	Good clinical annotation	Future higher grade within 1 year
	>50% cellularity	Squamous contamination
		Tumour contamination
		Absence of IM in non-progressors (i.e. only gastric metaplasia)

	Adjacent to cancer for the pre-cancer cohort

Table 3 Inclusion and exclusion criteria in cohort creation

IM = intestinal metaplasia

It was found, on reviewing the H&Es of cases, that the frozen sample rarely captured the highest grade which the patient was known to have at that timepoint, highlighting the problem of sampling bias. For example, a patient who had a small focus of HGD on their FFPE diagnostic biopsy, but the research frozen biopsy at that level was taken in an alternative quadrant and did not capture this focus. As a result, it was decided to accept any grade of dysplasia in dysplastic cases, rather than just the highest grade. E.g. a patient has IMC but the biopsy is composed of 50% HGD. Further rational for this was the difficulty for pathologists in assessing the frozen tissue, with the relative loss of structure due to the freezing process. The most difficult group to make was the LGD group, as many could not be used as contained only IM. The pre-progressor group was equally difficult to create, as it required a patient to have been surveilled for a significant number of years prior to progressing and had research biopsies on those occasions.

Multiple levels were later selected for six cases for the heterogeneity analysis. The same criteria as above were used for identifying these samples.

In total, 1161 frozen biopsies from 315 patients were found, cut and reviewed for inclusion. See Figure 4 for a summary of the cohort creation and Figure 5 for patient examples.

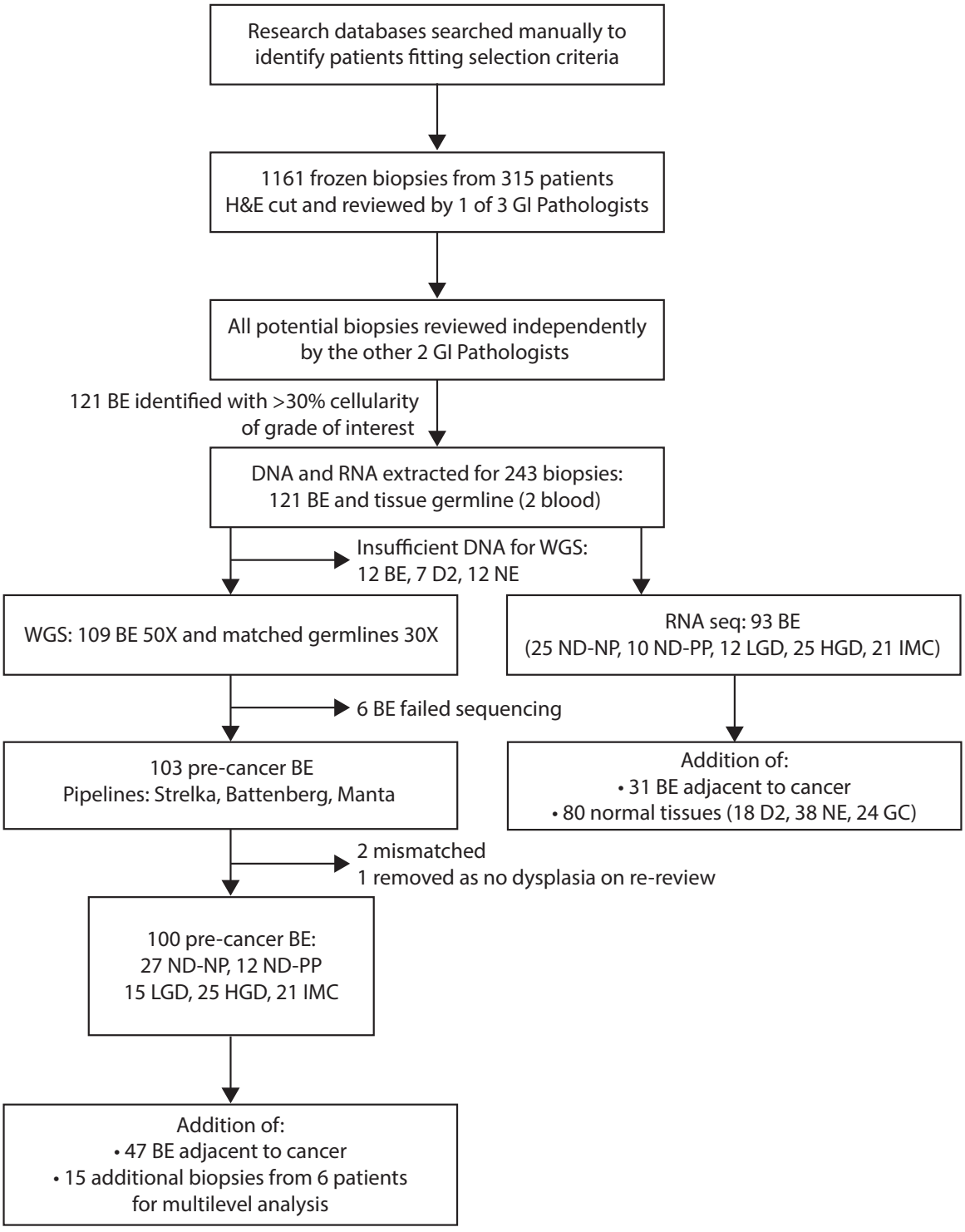


Figure 4 Flow diagram of cohort creation

GI = gastrointestinal, WGS = whole genome sequencing, BE = Barrett's oesophagus, ND = non-dysplastic, ND-NP = non-dysplastic non-progressor, ND-PP = non-dysplastic pre-progressor, LGD = low grade dysplasia, HGD = high-grade dysplasia, IMC = intramucosal carcinoma, D2 = 2nd part duodenum, NE = normal oesophagus, GC = gastric cardia

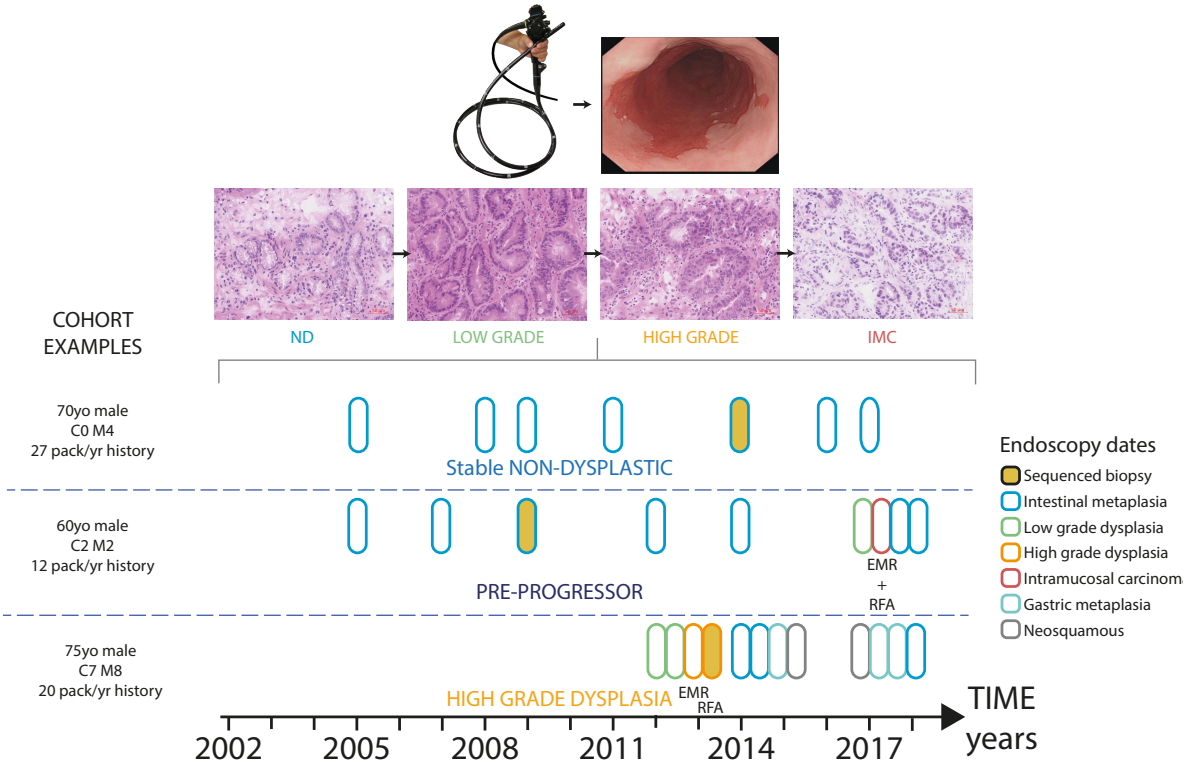


Figure 5 Cohort examples

Endoscopy, histology and timing of biopsies for example patients in cohort. H&Es of frozen biopsies shown to highlight the increased difficulty in pathology review due to the snap-freezing preservation technique. Timelines of patient progression and sample sequenced given. EMR = endoscopic mucosal resection, RFA = radiofrequency ablation, ND = non-dysplastic, IMC = intramucosal carcinoma. C and M values describe the circumferential (C) and maximal length (M) of the Barrett's oesophagus segment in cm.

2.2 DNA/RNA extraction

Whole frozen biopsies were homogenised on the Precellys® and DNA and RNA were extracted using the AllPrep DNA/RNA Mini Kit (Cat No. 80204; Qiagen®, Germany), as per protocol and performing all additional optional steps to maximise yield. DNA was eluted in 100ul EB buffer and RNA in 30ul RNA-free water. RNA was initially quantified using the Nanodrop. DNA was quantified using the Qubit® Low Sensitivity assay on the Qubit® 2.0 fluorometer (Invitrogen, Life Technologies, UK). 20ng/ul concentration of DNA was required for whole genome sequencing. For the pre-cancer cohort, DNA and RNA were extracted from 243 biopsies in total (121 BE and 122 germline) of which 31 had insufficient DNA for WGS (Figure 4). Blood was extracted as the germline reference for two cases using the QIAmp Blood Maxi kit (Cat No. 51192; Qiagen®, Germany). A further 62 biopsies and their matching normal tissue samples had DNA and RNA extracted for the multilevel and BE-adjacent to cancer analyses. A cohort of normal tissue biopsies (18 D2, 38 NE, 24 GC) had RNA extracted for comparison in the expression analysis. NE biopsies had an H&E cut prior to use to ensure they were only composed of squamous epithelium.

2.3 DNA library preparation and sequencing

A total of 124 pre-cancer BE biopsies (matched BE-germline) and 47 BE adjacent to cancer (Trio BE) were sequenced from 153 patients, under Illumina contracts. 100-bp paired-end sequencing was carried out to an average depth of 50x for BE 30x for matched normal.

Library preparation for the further multilevel samples was performed in house using the TruSeq DNA PCR-Free kit (Illumina, CA) as per protocol with a 2ug input. These 15 samples were run over 4 lanes of the NovaSeq 6000 sequencer (Illumina, CA) at the Cambridge Cancer Institute, Cambridge, UK, to a depth of 50x. Sequencing output from these batches were compared to previous sequencing to ensure that they were equivalent.

2.4 Whole genome sequencing analysis

2.4.1 Pipelines for variant callers

The FastQC package was used to assess the quality-score distribution of the sequencing reads and perform trimming if necessary. Read sequences were mapped to the human reference genome (GRCh37) using Burrows–Wheeler alignment (BWA-mem) 0.7.17 (Li

and Durbin, 2009). Duplicates were marked and discarded using Picard 2.9.5 (http://broadinstitute.github.io/picard/). Six BE samples failed WGS due to insufficient coverage. No samples had evidence of microsatellite instability using MSIsensor (Niu et al., 2014).

Somatic mutations and Indels were called using Strelka 2.0.15 (Saunders et al., 2012) with additional filters (Table 4).

Overall, 98% of the known genome was sequenced to at least 10x coverage and 60% to a 50x coverage. The whole cohort had at least 85% aligned bases within a read with a Phred quality of 20 or higher.

Filter	**Cut-off**
Variant Allele Count	< 4
Variant Allele Count Control	> 1
Distance to Alignment End Median	< 10.0
Distance to Alignment End MAD	< 3.0
Variant Map Qual Median	< 40.0
Map Qual Diff Median v	< -5.0 or > 5.0
Low Map Qual	> 0.1
Variant Base Qual Median	< 30.0
Variant Strand Bias	< 0.02
and Strand Bias	> 0.02
SNV Cluster 50	> 2
SNV Cluster 100	> 4
Repeat	>= 12

Table 4 Strelka filters

Structural variants were identified using Manta 0.27.2 (Chen et al., 2016). Discordant reads and split reads were used to identify putative breakpoint junctions. These methods have been compared to other variant callers in the ICGC benchmarking exercise and have among the best sensitivity and specificity (Alioto et al., 2015). Single nucleotide polymorphisms were called using GATK HaplotypeCaller 3.2-2 (McKenna et al., 2010). Copy number was initially called using ASCAT 2.3 (Van Loo et al., 2010); however, it was recalled using Battenberg v2.3.2 (Nik-Zainal et al., 2012a) which was able to call subclonal copy number for the clonality analysis.

Using the ASCAT output, WGD was seen in 11% ND-NP (3/27), 0% ND-PP, 21% LGD (3/14), 39% HGD (10/26), 30% IMC (6/20) (Figure 6). We were surprised by this calling of WGD in NP cases but further investigation did not highlight any obvious reasons for this. They were all *TP53* mutation negative and the cellularity estimates were not at either extreme (0.46, 0.51 and 0.23).

In comparison, with the Battenberg output all ND cases were diploid. The LGD and IMC proportions with WGD did not change from ASCAT, but the proportion of HGD cases with WGD halved to 19%.

Estimates of cellularity correlated well between ASCAT and Battenberg. However, a comparison of ploidy showed ASCAT to be possibly over-calling a number of samples (Figure 6). We looked at the raw ASCAT output for these cases and found that ASCAT had struggled to call these cases and the second alternative fit correlated better with Battenberg. Further comparisons between the two callers did not show any significant differences for calling clonal copy number so we decided to move over to this caller completely for the analysis.

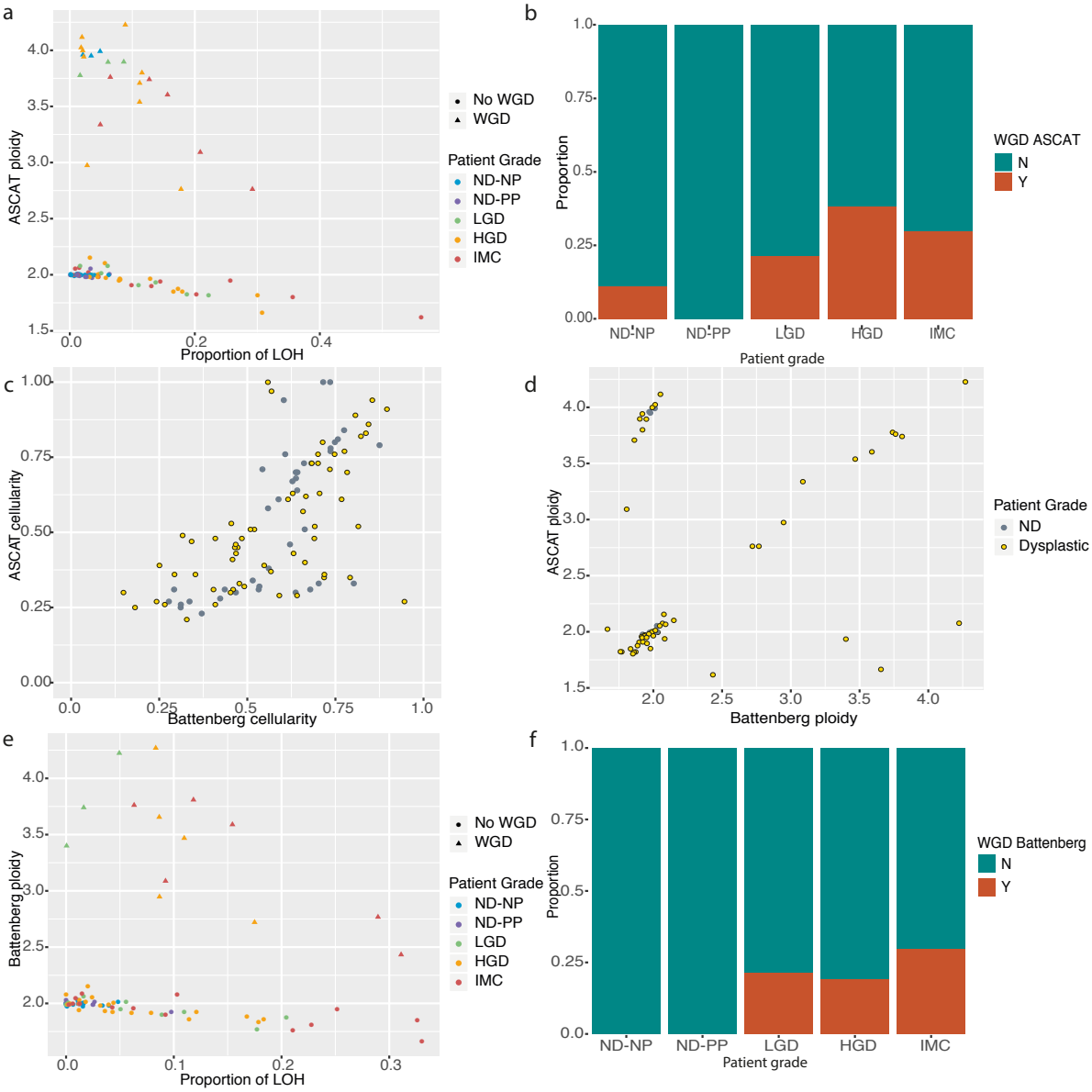


Figure 6 Whole genome duplication: ASCAT versus Battenberg

a. ASCAT ploidy estimates against loss of heterozygosity (LOH) proportion of genome. Samples in the upper cluster are considered to have undergone WGD. **b.** ASCAT proportions of samples with WGD per grade. **c.** Correlation between ASCAT and Battenberg cellularity output. **d.** Correlation between ASCAT and Battenberg ploidy estimates. **e.** Battenberg ploidy estimates against loss of heterozygosity proportion of genome. **f.** Battenberg proportions of samples with WGD per grade. WGD = whole genome doubling, ND = non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

DPClust was then run to cluster mutations by cancer cell fraction (CCF), given the copy number profile from Battenberg. After removal of artefactual mutation clusters containing <1% of the total number of mutations for that sample, if a clonal cluster of mutations was found (within CCF boundaries =0.95-1.05), and this had the highest CCF, or the largest number of mutations, this sample was classed as 'passed'. Where these conditions were not fulfilled, the sample was marked as 'failed' and required a rerun with new purity (rho) and ploidy (psi) parameters. Rho was calculated using the cluster that was closest to falling within the clonal boundaries. The rho value of the current call, and calculation of psi following that, was as follows:

rho_2 = rho_1*CCF of clonal cluster

psi_2 = ((rho_1 * psi_1) + 2*(rho_2 - rho_1))/rho_2

The copy number fitting steps of Battenberg and DPClust were then rerun to produce an updated call, and the test for 'passing' or 'failing' a profile was repeated. A large proportion (an additional ~20%) of samples which failed the first run typically now passed with these criteria.

For the small number which had not passed by this point, a further rerun was performed using a reference segment that was likely to have been called incorrectly by Battenberg. This was identified manually, and a new solution proposed, defined as a major and minor allele copy number. This proposed new copy number solution, along with information from calculation of the BAF and logR from the subclones file for this segment, was used to calculate a further estimate for the purity (rho) and ploidy (psi_t) of the sample.

The copy number fitting steps of Battenberg and DPClust were then rerun for a final a time to produce an updated call, and the test for 'passing' or 'failing' a profile was repeated, allowing for a further set of copy number profiles to be included into analysis.

Whole genome duplication (WGD) was called using the PCAWG method (Dentro et al.) which plots the ploidy relative to the fraction of the genome with loss of heterozygosity (LOH).

2.4.2 Mutational signatures

A de novo discovery of mutational signatures was performed using the non-negative matrix factorisation methodology (NMF) described by Alexandrov et al (Alexandrov et al., 2013) using the python version of SigProfiler (https://www.mathworks.com/matlabcentral/

fileexchange/38724-sigprofiler). For identifying optimal de novo signatures, we ran NMF for 2-10 ranks for 1000 iterations. This process identified 3 optimal signatures which, when decomposed, mapped to 8 known signatures. The de novo signatures were compared to the 50 known published COSMIC signatures (https://cancer.sanger.ac.uk/cosmic/signatures/SBS/).

2.4.3 Structural variant signatures

SV signatures were identified based on a framework as explained in Nik-Zainal et al., 2016. Briefly, SVs were classified into 38 categories based on the type and size of SV event, then were further classified into clustered and non-clustered. Events were considered to be clustered in a sample if a region of the genome (1Mb) was covered by >10 breakpoints. NMF was then applied on these events using Palimpsest (Letouzé et al., 2017), identifying 5 optimal signatures.

2.4.4 Chromothripsis

We identified both low and high confidence chromothripsis event based on oscillating copy number events in regions with clustered breakpoints across all samples using ShatterSheek (Cortés-Ciriano et al., 2018).

2.4.5 Kataegis

Clustered mutations representing Kataegis-like events within small genomic loci (5kb) were identified across all samples using a package in R called ClusteredMutations (https://cran.rproject.org/web/packages/ClusteredMutations/index.html).

2.5 RNA sequencing

2.5.1 Library preparation

RNA was quantified using the Qubit High Sensitivity RNA kit (Thermo Fisher) and checked for quality (RNA integrity number; RIN) on the Agilent 2100 Bioanalyzer® (Agilent Technologies, USA) using the RNA 6000 Nano kit. Samples with insufficient material, or an incalculable RIN were excluded. There was no other lower limit for RIN inclusion.

Samples were randomised to 3 batches, ensuring an equal spread of RIN values across the batches. Libraries were prepared with an input of 150ng RNA using the TruSeq Stranded Total RNA High Sensitivity protocol with ribosomal depletion. Samples with less than the specified input, but with >100ng total were included and this was noted for the analysis. Libraries were validated using the Agilent 2100 Bioanalyzer with the DNA 1000 kit and KAPA quantification (KAPA Biosystems, Roche, Switzerland) and were pooled according to the Illumina protocol. Samples were run on the HiSeq 4000 instrument to generate 75bp paired-end reads. A mixture of normal expression controls was run on each plate: squamous oesophagus, gastric cardia, duodenum. Duodenum shares some features of intestinal appearance of BE and it is hypothesized that BE arises from gastric cells. Squamous oesophagus is a less useful comparison because it shares few features with the glandular epithelium of BE.

2.5.2 Pipelines for RNA

RNA sequencing data was trimmed for poor quality bases using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and was then aligned using STAR (Dobin et al., 2013) using ENSEMBL gene annotation. Reads per gene were quantified using the summariseOverlaps function from the GenomicRanges package, which was also later used for computing Transcripts per million (TPM). Normalised expression data was corrected for batch effect using the ComBat function in the sva package (v3.20.0) (Johnson et al., 2007). Expression was calculated as log2(1+TPM). Principal component analysis was performed selecting the 1000 most variable protein-coding genes. DESeq2 software (v1.18.1) (Love et al., 2014a) was used to compare the differential expression between the different groups using the raw counts. Genes were considered to be significant if there was a >3-fold change and $p < 0.05$. Pathway analyses were performed using DAVID

6.8 (Huang et al., 2009b, 2009a), Ingenuity Pathway Analysis (IPA; Qiagen, Germany), StringDB (https://string-db.org) and Gene Set Enrichment Analysis.

2.5.3 Copy number driver gene discovery

GISTIC 2.0 (Mermel et al., 2011) was used to identify recurrently amplified and deleted regions from the raw copy number calls; a method that has been previously used in the lab (Frankell et al., 2019). Peaks were widened by 1 million base pairs up- and down-stream and all genes falling within these regions were considered. The expression (calculated as the log2(1+TPM) of each gene from the RNA seq data was compared for high vs. normal CN samples for each gene and mean expression levels compared using the Wilcoxon rank sum test. The Benjamini-Hochberg method was used to correct for multiple testing.

2.5.4 Immune signatures and chromosomal instability

Markers for both immune cell types and chromosomal instability were retrieved from publication. (Carter et al., 2006; Tamborero et al., 2018) and Gene Set Variation Analysis (GSVA) was used to assign enrichment scores to samples based on the expression of different markers in bulk RNA seq (Hänzelmann et al., 2013).

2.6 Heterogeneity/clonality methods

DPClust v2.2.5 (Nik-Zainal et al., 2012a) was used to model clonal expansions by calculating the cancer cell fraction (CCF) of each mutations (as described above). Clusters containing fewer that 1% of the total number of mutations were excluded. Phylogenetic trees were constructed manually using this output.

2.7 Clinical modelling

The rpart package in r was used with 16 features from genomic, expression and clinical data to grow a classification tree.

3. Results 1: The genomic landscape of Barrett's oesophagus

Aim 1: Elucidate the key biological processes driving Barrett's oesophagus (BE) to progress to oesophageal adenocarcinoma (OAC) by performing an integrated analysis of genomic and transcriptomic sequencing of the individual grades of BE.

- Perform a genomic characterisation of the grades of BE.
- Compare non-dysplastic samples that do not progress to those prior to progressing to see if genomic alterations can be defined which predict that progression will occur.
- Compare the genomic alterations in BE sampled from adjacent to cancer to pre-cancer BE.

3.1 Cohort selection and demographics

In order to characterise the stages of Barrett's oesophagus (BE) progression, samples were identified from patients with different grades of disease from non-dysplastic (ND) BE to intramucosal carcinoma. ND samples were split into two categories: samples from long-term ND patients who had never gone on to progress (denoted non-progressor, NP); and patients who had gone on in the future to progress to dysplasia (denoted pre-progressor, PP). Dysplastic samples were split into low grade (LGD), high grade (HGD) and intramucosal carcinoma (IMC). Strict criteria were applied for sample selection and these criteria, plus further details for cohort creation, are detailed in the Methods. All biopsies in the cohort were reviewed independently by three specialist BE pathologists. Biopsies were categorised both by the highest grade which the patient had reached on the date of the endoscopy and by the highest grade seen in the H&E of the sequenced biopsy.

In total, 1161 frozen biopsies from 315 patients were identified, cut and reviewed in order to create the final pre-cancer cohort (Figure 7). All of these cases had germline samples available to determine somatic mutations from inherited polymorphisms. Blood had not been taken in the study, so we used duodenum, in preference to normal oesophagus. This is because it is far removed from the BE and not at risk of being genetically altered by a field effect. Previous studies in the literature have predominantly used BE adjacent to adenocarcinoma to elucidate the progression stages (Ross-Innes et al., 2015a; Stachler et al., 2015). Therefore, as a comparison to our cross-sectional cohort of pre-cancerous disease stages, we also included samples from BE adjacent to cancer from 47 patients: so-called Barrett's oesophagus Trios (BE, cancer, germline).

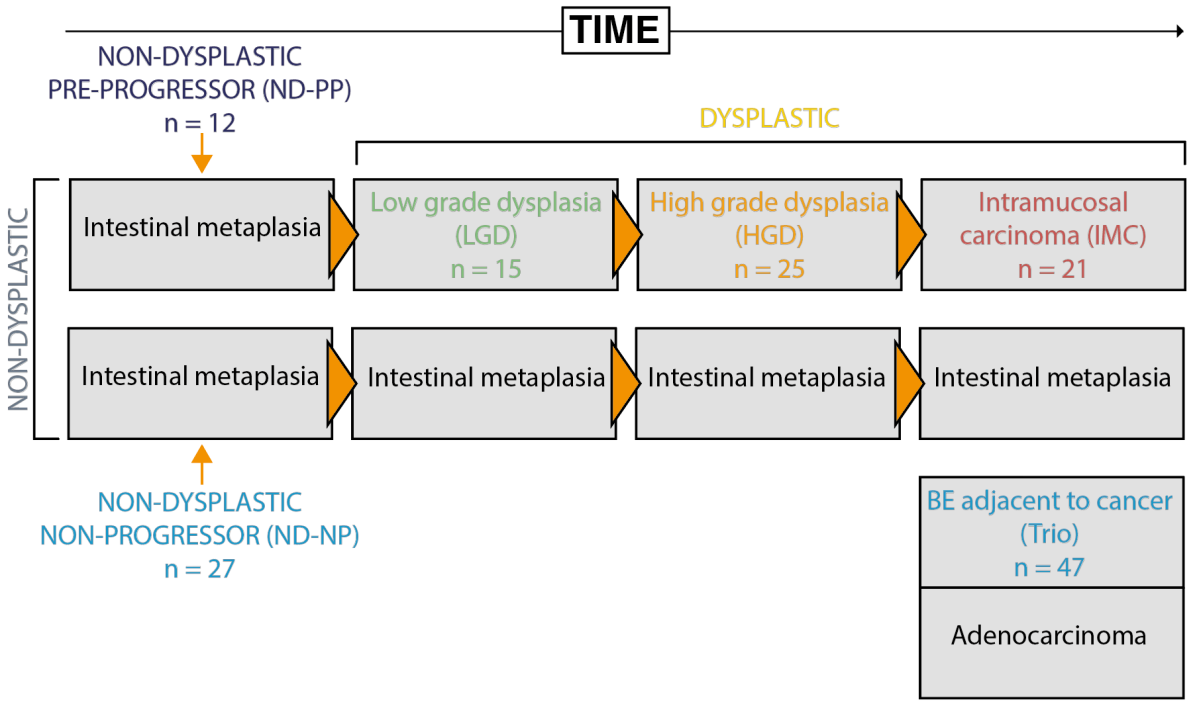


Figure 7 Cohort design

Cohort design and annotation terminology

The final cohort was determined by which cases had successful generation of whole genome sequencing (WGS) data. Six samples failed QC at Illumina and were excluded. Two cases appeared hypermutated, but further investigation showed this to be due to a mismatch between the sample and the germline reference: one ND-NP with 231,692 mutations; and a ND-PP with 239,153 mutations. One pre-progressor sample was excluded from further analysis because, on pathology consensus review, the criteria for inclusion had not been met. It had been taken at a LGD time-point only two months prior to a further endoscopy which found HGD. The biopsy available for sequencing contained only IM, and so was not suitable for inclusion in the cohort. This resulted in a pre-cancer BE cohort of 100 patients (27 ND-NP, 12 ND-PP, 15 LGD, 25 HGD and 21 IMC) plus 47 Trio BE (Table 5). The median total follow-up in the ND-NP was 138 months (range 45-251) in order to give confidence that they were long-term non-progressors.

Male gender, increasing age, length of the BE segment and smoking are recognised risk factors for the development of BE (Coleman et al., 2014; Krishnamoorthi et al., 2018). However, there were no significant differences between the three groups for gender or age

(complete data available for all patients); or length of BE between ND and dysplastic (complete data available for all pre-cancer cohort patients but not recorded for the Trios BE) (Table 5). There was a significant increase in smoking status (complete data for 80% of cohort) in the dysplastic group versus ND (ND 51.7%, dysplastic 83.0%; p value = 0.0047, Fisher's Exact Test) but the increase in the Trio BE cases compared to ND was not significant (p value = 0.13). The use of a proton pump inhibitor (PPI) has also been found to lower the risk of progression (Krishnamoorthi et al., 2018).We saw a trend towards fewer Trio BE patients being on a PPI (ND 94.9%, dysplastic 96.7%, Trio BE 77.5%, p value = 0.04 Fisher's Exact Test; 97% complete data). There was no significant difference between BMI (89% complete data) or non-steroidal anti-inflammatory (NSAID; 77% complete data) use. The cohort was predominantly of white ethnicity. This may be reflective of the higher incidence of BE amongst Caucasians (Corley et al., 2009), but also the demographic of the East of England region in which this study took place.

	Non-dysplastic	Dysplastic	BE adjacent to cancer (Trio)	p value ND vs. Dysp	p value ND vs. Trio
Number of patients	39	61	47		
Grade of patient at time	NP: 27, PP: 12	LGD: 15, HGD: 25, IMC: 21	OAC: 47		
Highest grade in frozen biopsy	ND: 38, LGD: 1	LGD: 28, HGD: 28, IMC: 5	ND:38, Indefinite:1, LGD: 5, HGD:3		
Gender, M:F	3.3:1	4.5:1	6.8:1	0.61	0.26
Male	30	50	41		
Female	9	11	6		
Age in years, median (range)	69 (31-83)	69 (39-86)	67 (49-84)	0.62	0.48
BMI, median (range)	28.8 (23.2-43.9)	28.8 (13.5-40.9)	28.1 (20.3-38.8)	0.96	0.5
NA	5	2	9		
Maximum length BE segment in cm, median (range)	5 (2-14)	6 (0.5-18)	Not recorded	0.54	
Ethnicity					
White British	35	34	36		
White other	3	1	1		
Pakistani	0	1	0		
NA	1	5	10		
Smoker					
Y, n (%)	15 (51.7%)	39 (83.0%)	30 (71.4%)	0.005	0.13
N	14	8	12		
NA	10	14	5		
PPI					
Y, n (%)	37 (94.9%)	58 (96.7)	31 (77.5%)	0.65	0.04
N	2	2	9		
NA	0	1	7		
NSAID					
Y	11 (29.7%)	20 (33.3%)	6 (37.5%)	0.82	0.75
N	26	40	10		
NA	2	1	31		
		LGD:6 (0-88)			
Surveillance pre study sample in months, median (range)	NP: 108 (5-227) PP: 48 (0-141)	HGD: 2 (0-193) IMC: 3 (0-168) LGD: 65 (0-83)	NA		
Follow-up post study sample in months, median (range)	NP: 38.5 (0-116) PP: 92 (26-119)	HGD: 26 (0-127) IMC: 3 (0-168)	NA		

Table 5 Clinical demographics of the final cohort

P values calculated using Fisher's Exact test for categorical variables and T-test for continuous variables. Statistical comparisons made between ND and dysplastic groups, and ND and Trio groups. BMI = Body mass index, PPI = proton pump inhibitor, NSAID = Non-steroidal anti-inflammatory drugs, BE = Barrett's oesophagus, ND = non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma, OAC = oesophageal adenocarcinoma. % given exclude NA patients.

3.2 Pre-cancer Barrett's oesophagus cohort

3.2.1 Mutational burden

Next we considered the overall genomic features of the groups, starting with the mutation burden. BE biopsies were found to be highly mutated, as previously described (Ross-Innes et al., 2015a; Stachler et al., 2015).There was a median of 14,126 mutations (IQR 9,580-20,914) across the coding and non-coding genome, with a median mutation burden of 4.5 mut/Mb. None of the samples had microsatellite instability. All sequencing studies conducted on tissue samples can be affected by the cellularity of the sample i.e. the proportion of BE cells within the whole sample. Large proportions of normal tissue will lower the allele frequency of a mutation and may take it below the threshold for being called by Strelka. However, the mutation burden did not generally correlate with the cellularity of the biopsy calculated by the Battenberg algorithm (Figure 8a) One LGD sample (LP6008280-DNA_B04) had only 175 mutations. Battenberg had difficulty calling the cellularity of this sample and so the low mutation burden was likely false and due to a low cellularity of BE in the biopsy. It was excluded from further analysis (circled in red in Figure 8a).

The median mutation burden increased linearly with patient grade, but with wide ranges (Figure 8b). There was a significant difference in mutation burden when LGD, HGD and IMC were grouped together as dysplastic and compared with all ND cases (median ND 3.19 mut/Mb; median dysplastic 5.62mut/Mb; $p = 5.9x10^{-6}$, Wilcoxon Rank Sum test) (Figure 8c). There was no evidence that the actual grade within the frozen biopsy correlated better with mutation burden than the highest grade of disease in the patient, suggesting that the overall clinical patient status was most informative (Figure 8b). This may be because there is a local effect within the segment where the dysplasia arises. However, it may also be because of the difficulty for pathologists in distinguishing between LGD and HGD in frozen biopsies because of the artefact caused by ice crystal formation and thawing. This was exemplified by 7 of the 14 biopsies which were from patients diagnosed with LGD from their surveillance FFPE biopsies, however the pathology review consensus of the frozen was of HGD. These biopsies were analysed using the diagnostic FFPE pathology grades for this reason.

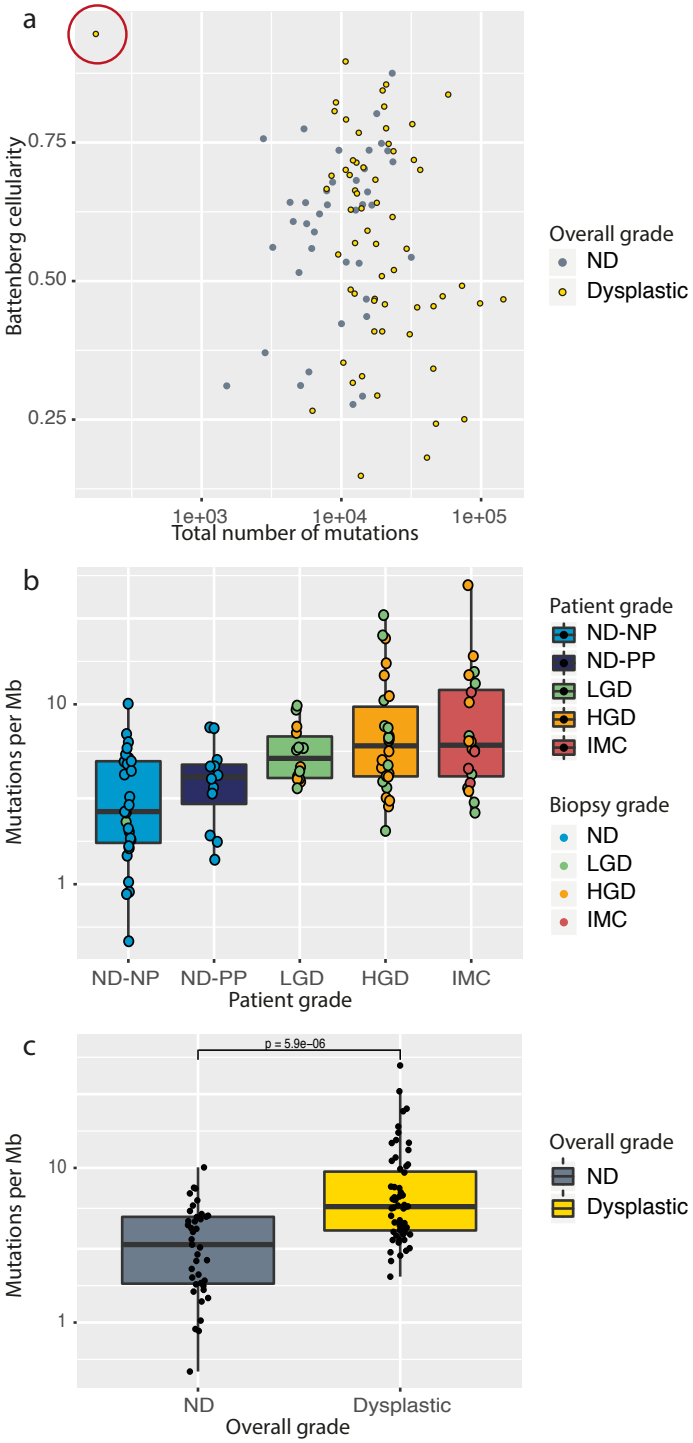


Figure 8 Mutation burden across the grades

a. Scatterplot showing the relationship between the cellularity called by the Battenberg algorithm and the total number of mutations per sample. The sample circled in red were excluded from subsequent analyses because of a low mutation count due to low cellularity. **b.** Mutation burden per grade of patient that biopsy taken from. Dots coloured by highest grade of Barrett's oesophagus (BE) within the biopsy: which may be lower than the highest grade in the patient overall. **c.** Mutation burden in ND versus dysplastic BE. y axis log10 scale. ND = Non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma. P values calculated using Wilcoxon Rank Sum test.

Next, we determined the mutational signatures in the samples. Mutational signatures consider the proportions of each base substitution within the context of the immediately 5' and 3' bases. We applied the algorithms derived by Alexandrov et al (Alexandrov et al., 2013) using SigProfiler (Alexandrov, 2019). A de novo discovery of mutational signatures using non-negative matrix factorisation extracted 9 signatures (Figure 9). SigProfiler automatically compares the de novo signatures to the known mutational signatures (https://cancer.sanger.ac.uk/cosmic/signatures/SBS/) in order to assign their likely aetiology, if known. The proportions of these signatures in each sample are shown in Figure 10. No new signatures were discovered but this was not unexpected given the arguably small size of the cohort for a de novo discovery.

All of the 9 signatures, except signature 9, are seen in OAC, particularly signatures 1 (ageing; C>G substitutions in a *CG context) and 5 (unknown aetiology; T>C in an AT* context). Signatures 17a (T>C in a CTT context) and b (T>G in a CTT context) are considered the hallmarks of OAC (Secrier et al., 2016). They are closely related signatures of unknown aetiology, although thought to be caused by reflux/reactive oxygen species (Pich et al., 2019; Tomkova et al., 2018). They are commonly found in the same samples and have previously been shown to correlate with mutational burden in OAC (Secrier et al., 2016) which was also the case in our pre-cancer cohort (Figure 11). The mutational signature profile has previously been shown to be unchanged between the cancer and its adjacent prevalent BE (Ross-Innes et al., 2015a) but it has not been known when it is set. Here we found that signature proportions did not change significantly across the pre-progression grades. Signatures 17a and b were present even in the ND samples i.e. from an early stage. Signature 18, observed in 20% of OAC, was also present from the early ND stage. We also observed signature 1, associated with ageing, to negatively correlate with mutational burden. Although the total proportions of the signature (maximum 20% total proportion) was lower than those of signature 17.

Signature 2 (APOBEC-driven hypermutation; (Nik-Zainal et al., 2012b)) was seen in 52% of the cohort and is in 39.5% of OAC. Signature 9, which is not seen in OAC, was present in 3 samples (3%). We did not observe representation of signature 3 (BRCA-related; (Nik-Zainal et al., 2012b)) which is present in 7% of OAC (Alexandrov et al., 2018).

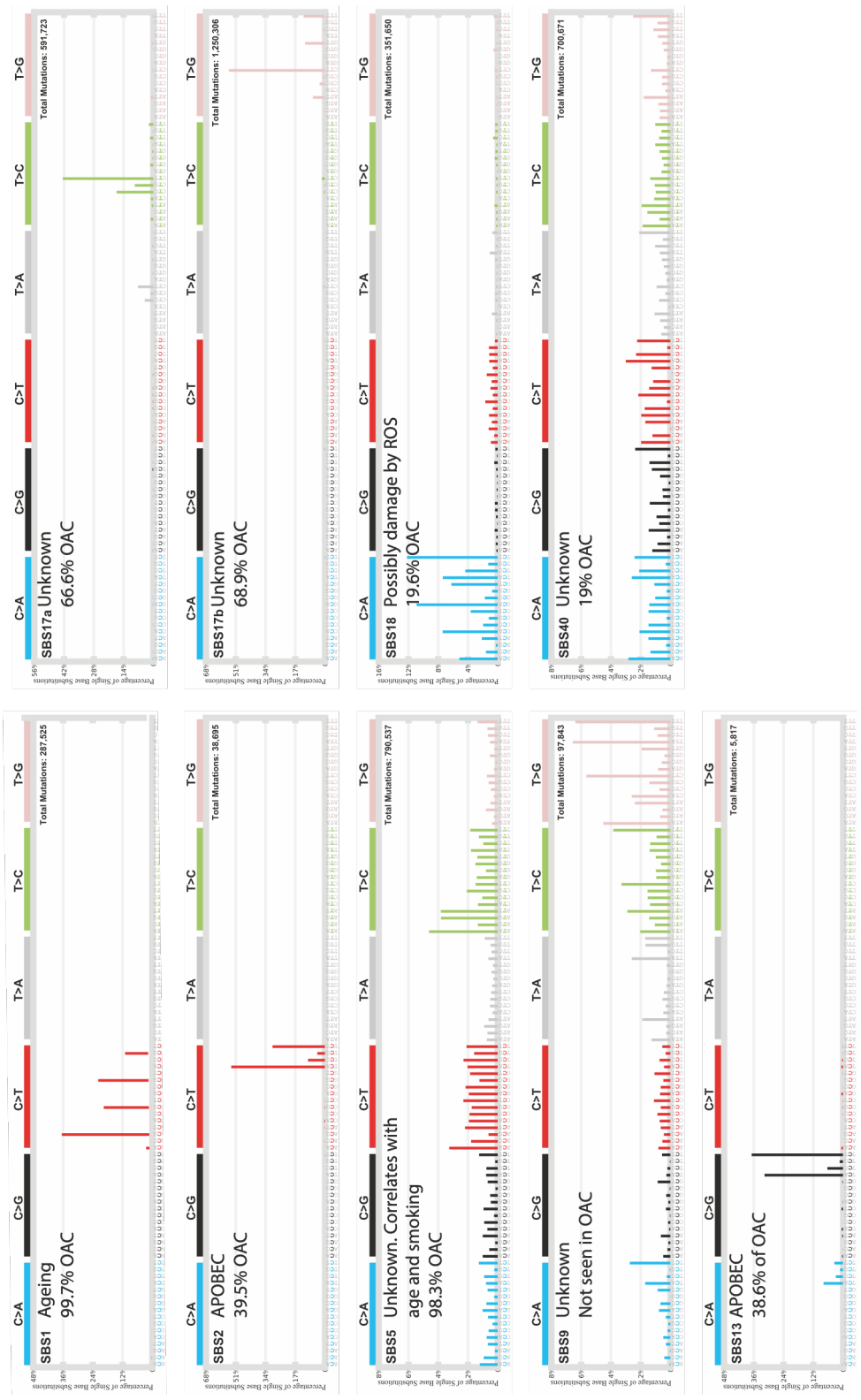


Figure 9 De novo discovery of mutational signatures in the cohort

Nine signatures found corresponding to known COSMIC database signatures. Percentage of each single base substitution (SBS) drawn for each signature. Signature aetiology identified, if known, and percentage of oesophageal adenocarcinomas (OAC) with each signature (taken from Alexandrov et al. 2018, which used TCGA data). ROS = reactive oxygen species.

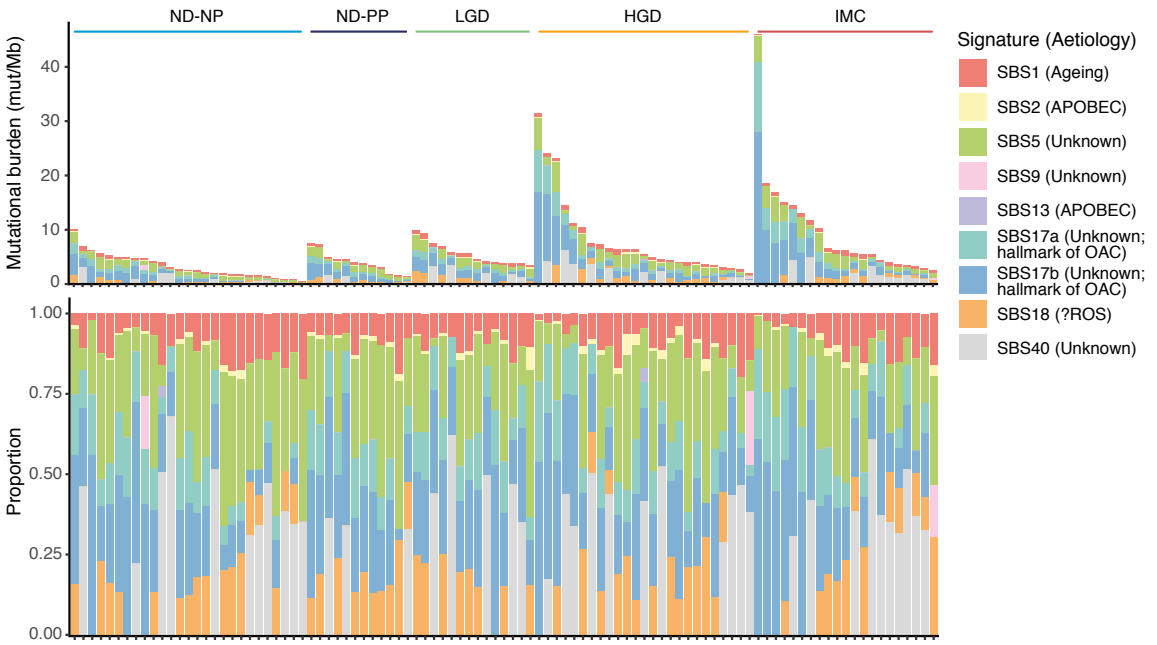


Figure 10 Mutational signatures in the cohort

Number of mutations per case contributing to each signature and proportions of each signature per case. Grouped by grade and ordered within that from high to low mutation burden. SBS = single base substitution, ND = Non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma, ROS = reactive oxygen species, OAC = oesophageal adenocarcinoma.

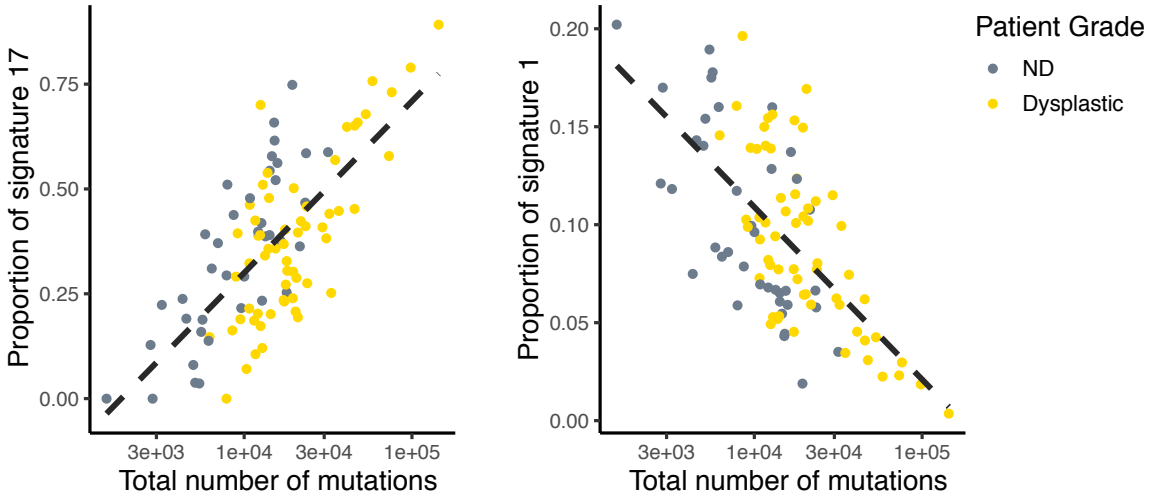


Figure 11 Correlation of mutational signatures 17 and 1 with mutation burden

Proportion of signature 17 (A+B) or signature 1 in each sample plotted against mutation burden. Dots coloured by grade of patient. Dashed line is line of fit by linear regression. ND = non-dysplastic.

3.2.2 Copy number aberrations across the grades

In view of the importance of copy number alterations in OAC, this was evaluated in the pre-cancerous BE stages. The Battenberg algorithm (Nik-Zainal et al., 2012a) was used to call copy number, tumour purity and ploidy. Battenberg was chosen instead of ASCAT because it is able to estimate both clonal and subclonal events. A comparison of these two methods for our dataset is included in Methods.

The mean ploidy of the ND was 2.00 (median 2.00, range 1.92-2.02) and the dysplastic mean 2.29 but with a wider range up to 4.27 (median 1.99, range 1.67-4.27) highlighting the likely presence of duplicated genomes amongst the dysplastic cases. From the Battenberg output we were able to distinguish the types of CNAs that were occurring: amplifications, amplifications with loss of heterozygosity (LOH) of the minor allele, deletions, deletions with LOH of the minor allele and copy number neutral LOH. In this chapter, total % CNA refers to the addition of the proportions of these 5 types of gain and loss multiplied by 100. Figure 12a shows the proportions of each of these alterations within the samples. Copy number neutral LOH was the most frequent alteration observed, followed by deletion LOH. As seen with mutation burden, dysplastic cases had significantly more CNAs than ND cases (median ND 7.9%, dysplastic 11.4%, p value = 1.5×10^{-6}, Wilcoxon Rank Sum test), with a wide variance between cases which increased with grade of dysplasia (Figure 12b). Of the five most extreme dysplastic outlier cases, circled in red in Figure 12c, four were mutant for *TP53*. There was one ND-PP with a higher % CNA than the other ND samples but this case was also not *TP53* mutant. However, what was particularly noticeable in these five cases was the high number of structural variants (SV): median 292 total SV count, range 128-517). This will be discussed further in the next section.

When we performed a sub-analysis by the grade of the patient there was a trend towards an increase in copy number (Figure 12c). However, there was no correlation within a sample between its total mutation burden (TMB) and CNA (Figure 12d), such that a high TMB was occurring in different samples to those with a high percentage CNA. We also considered whether the grade composition of the frozen biopsy, rather than the grade of the patient, was affecting the CNA levels seen but this did not appear to be the case. A LGD biopsy from a HGD patient had the highest CNA in the HGD patient group, as did a HGD biopsy in the IMC patient group (Figure 12c).

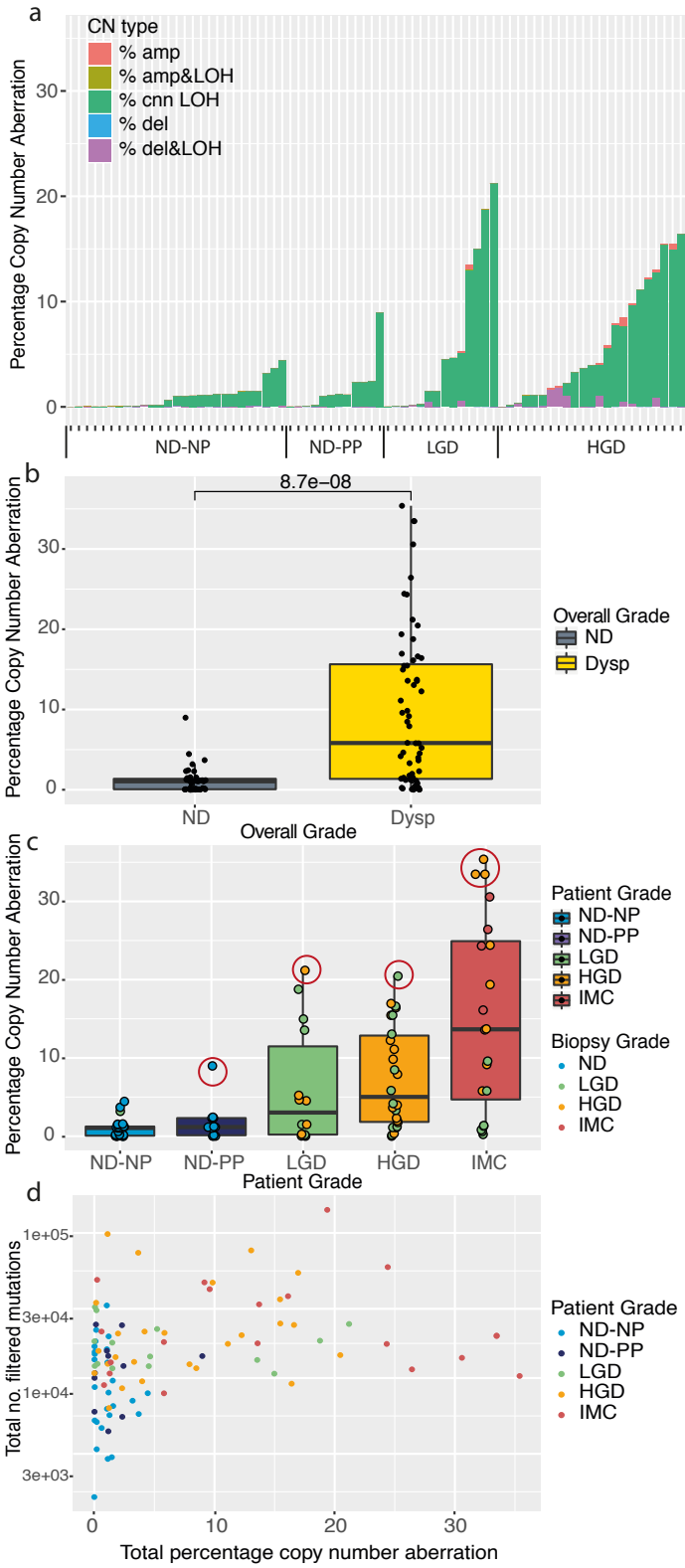


Figure 12 Copy number aberrations across the grades

a. Percentage and type of copy number aberration per sample ordered by grade. **b.** Copy number aberration in non-dysplastic versus dysplastic samples. **c.** Copy number aberration per grade of patient that biopsy taken from. Dots coloured by highest grade of BE within the biopsy. Dots circled in red are outliers discussed further in the text. **d.** Comparison of copy number and mutation burden per sample. ND = Non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma. P value calculated using Wilcoxon Rank Sum test.

Given that whole genome duplication (WGD) is a common event in OAC we wanted to look at the proportions of samples with WGD and the timing of this occurrence. The PCAWG method (Dentro et al.) was used to determine if a sample had undergone WGD. The method plots the ploidy relative to the fraction of the genome with loss of heterozygosity (LOH). Using the Battenberg output, WGD was not observed in any of the ND-NP or ND-PP (Figure 13). LGD and HGD had similar percentages 21% (3/14) and 19% (5/26) respectively. This increased to 30% in IMC (6/20) but the increase was not significant compared to HGD (p value = 0.52, Fisher's Exact test). We analysed the cancers matching the BE Trio samples as a comparison and observed WGD in 75%. Other studies have observed 62.5% (Stachler et al., 2015) and 50% (Secrier et al., 2016) WGD in OAC, but used different methods for calculating it e.g. ploidy > 2.7. When we used this cruder cut-off, 59.6% of the tumours were whole genome duplicated. So, a similar percentage to the other studies.

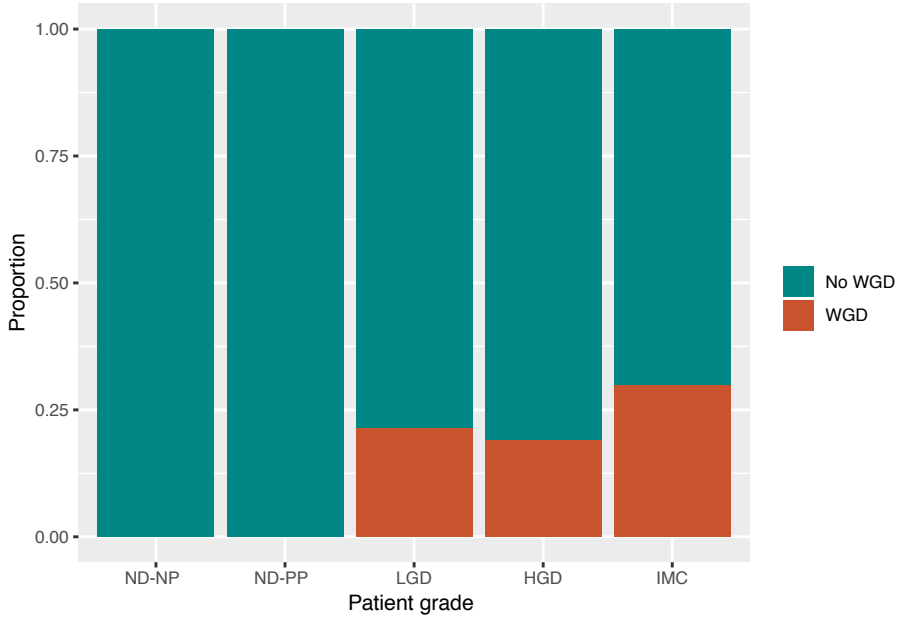


Figure 13 Whole genome duplication in the cohort

Battenberg estimates of whole genome duplication (WGD). ND = Non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

We then used the WGD as a reference to determine the relative timing of other gains and losses. This analysis was performed by considering the variant allele frequencies of mutations in relation to the copy number in that region and the sample purity, to give the number of chromosomes with the mutation. From this, mutations were split into those that occurred before the chromosomal gain (as they are in two alleles) or afterwards (in only one allele).

Figure 14 plots the timing of the CN gain relative to the mutational time. 100% indicates when all the mutations in a specific sample had occurred. Timing can only be relative, not real-time, given that mutation acquisition occurs at different rates in different patients. We found in our whole genome duplicated samples that there was a prolonged mutational period prior to the duplication. In most cases more than 80% of mutations had been acquired prior to the duplication; indicating that WGD is a late event. However, in the non-WGD samples the copy number gains were more evenly acquired over the duration of mutational time. WGD has previously been suggested to be an early event, with subsequent oncogene amplification and rapid progression to OAC (Stachler et al., 2015). However, they focussed on OAC and its adjacent BE. It is possible that we are capturing an alternative evolutionary pathway by studying samples prior to progression.

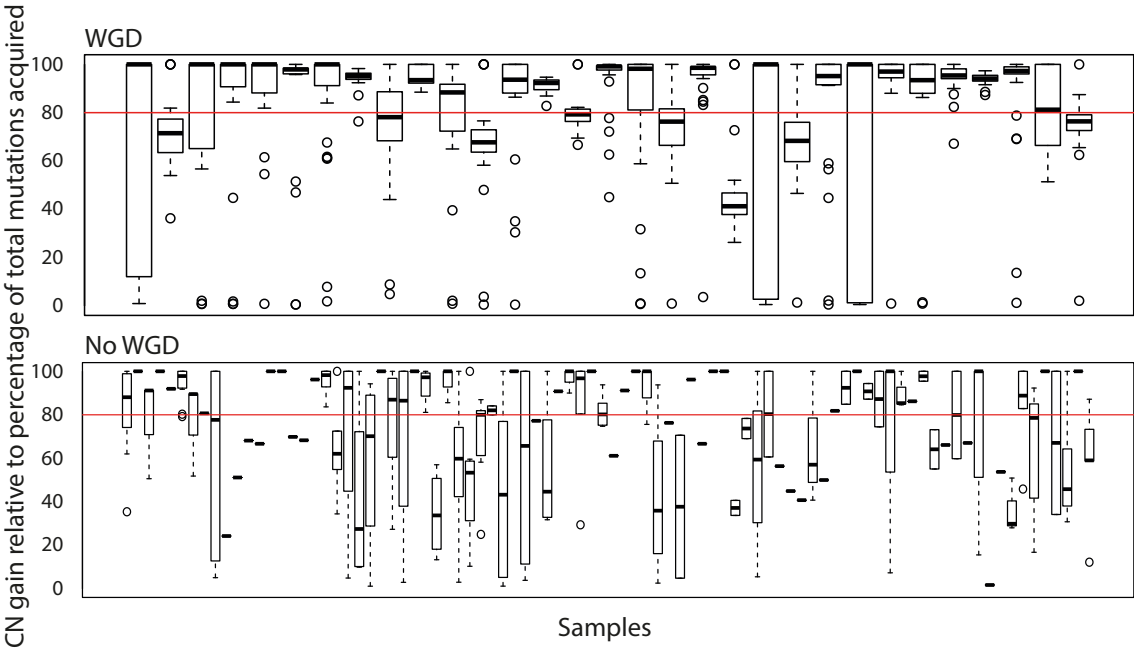


Figure 14 Timing of whole genome duplication

The y axis represents a timeline from the first mutation to the acquisition of 100% of mutations in an individual sample. The box plots mark the percentage of mutations present when each copy number (CN) gain in the sample took place. The red line is an arbitrary threshold at 80% which shows that most of the CN gains in whole genome duplicated (WGD) cases occurred after this threshold. Whereas samples with no WGD gained CN in a more evenly distributed manner, over the lifetime of mutation acquisition.

We next looked at the locations and sizes of these CNAs across the genome in order to understand if specific regions were recurrently affected. Figure 15 shows the amplifications and deletions per chromosome for each patient. Regions were binned according to the number of breakpoints in the region, across the whole cohort. The genomic landscape was dominated by deletions. The most commonly deleted loci were the fragile site in chromosome (chr) 3 containing *FHIT* and the locus of chr 9 containing *CDKN2A*.

Hierarchical clustering divided the cohort into two main clusters; however, there was no clear distinction between ND and dysplastic samples. Earlier in the analysis we had seen a significant difference between ND and dysplastic cases when just considering the means of the total proportions of CNAs (Figure 12). But the ranges had been wide and a number of dysplastic samples had had low proportions of CNAs. In this regional analysis we did not see distinct genomic loci which cleanly split the grades.

One main branch of the clustering dendrogram predominantly consisted of dysplastic samples. However, two distinct groups of dysplastic samples, with fewer CNAs, clustered with ND. Both these groups were lacking the large chr 9 deletion, which dominated the other dysplastic cases. One cluster did not have a deletion of the short arm of chr17 (containing *TP53*) but instead amplifications of parts of chr 8, 9, 17 and 18. The other cluster was dominated by both amplifications and deletions in chr 16-22. The samples in these clusters were a mix of LGD to IMC. Hence, the clustering by copy number was, to some extent, independent of the grade of dysplasia. Overall, from this analysis, the transition from ND to the extremes of dysplasia seems to be more of a continuum, with a gradual acquisition of copy number alterations; some of which are hotspots e.g. the fragile sites, and others private to the sample.

We considered individually the 13 dysplastic outlier cases with fewer CNAs, clustering more with the ND samples in Figure 15 (marked as dysplastic outliers). Whilst they had fewer deletions, amplifications within chr 8, 9, and 18 were prevalent. These amplifications will be further considered later in the chapter with the driver analysis. Seven of the cases had clear alternative explanations for progression: 2 were whole genome duplicated, 4 were *TP53* mutant and the others either had high mutation burdens or numbers of SVs despite little copy number change (Table 6). None exhibited chromothripsis.

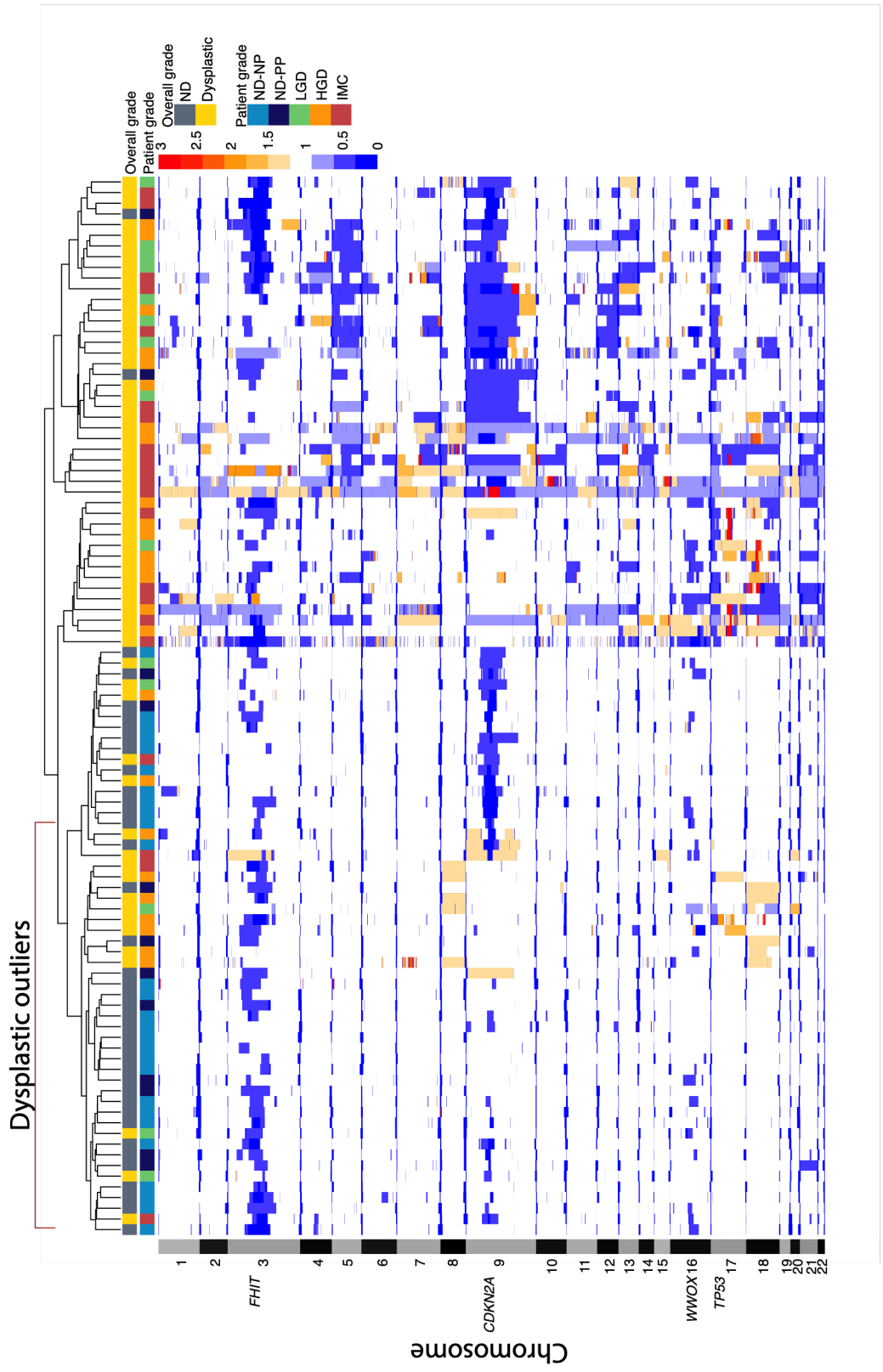


Figure 15 Hierarchical clustering of amplifications and deletions by locus

Heatmap of patients (x axis) and genome by chromosome (y axis). Chromosomes binned based on number of breakpoints across the cohort. Amplifications represented in red and deletions in blue. Hierarchical clustering using Ward D segregates dysplastic (yellow) and non-dysplastic (grey). Deletion across all samples in chromosome 3 represents the fragile site containing *FHIT*. Similarly, *WWOX* is found in the deleted fragile site on chromosome 16. *CDKN2A* and *TP53* lie within the deletions in chromosomes 9 and 17 respectively. ND = Non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

Position in clustering	Patient Grade	Age	Gender	Purity	TMB (mut/Mb)	SV count	Total % CN Change	*TP53* mutation	WGD	Ever Smoked
2	IMC	64	M	0.85	6.67	372	0.60	0	0	Y
6	LGD	67	M	0.72	3.86	52	0.04	0	0	Y
10	LGD	73	M	0.64	5.72	136	0.02	0	0	Y
26	HGD	86	F	0.63	3.73	305	8.49	1	1	Unknown
27	HGD	73	F	0.59	4.90	57	0.35	1	0	Y
29	HGD	62	M	0.46	31.35	144	1.10	0	0	Unknown
30	HGD	71	M	0.78	6.66	147	4.18	0	0	Y
31	LGD	69	M	0.32	3.85	56	4.54	1	1	Y
32	HGD	67	F	0.79	3.44	51	0.01	0	0	Y
34	HGD	60	M	0.69	2.71	54	2.23	1	0	Y
35	IMC	75	M	0.81	2.85	105	0.81	0	0	Unknown
36	IMC	81	M	0.71	4.09	80	1.36	0	0	Unknown
38	HGD	75	M	0.66	4.00	62	1.11	0	0	Y

Table 6 Genomic and clinical features of the outlier dysplastic cases on hierarchical clustering of copy number aberrations

Values highlighted in orange indicate possible alternative reasons for progression despite clustering with ND. TMB = total mutation burden, SV = structural variation, CN = copy number, WGD = whole genome duplication, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

3.2.3 Structural variation

In order to call structural variants (SVs: translocations, inversions, large scale deletions and duplications) we used the Manta pipeline (Chen et al., 2016). When the total number of SVs in each sample was calculated we found that there was a larger and more significant difference between ND and dysplastic samples that out-weighed the variables considered so far: ND median 43 (IQR 35.0-60.5), dysplastic median 140.5 (IQR 78.5-256), Wilcoxon Rank test p value = $1.6x10^{-11}$ (Figure 16a). It was also observed that the ranges of SV count per sample was wider for the HGD (22-876) and IMC (38-1602) samples versus the ND-NP (9-85) and ND-PP (10-128) (Figure 16b). Of note, total numbers of SVs did not correlate with either CNAs or ploidy on a per-sample analysis (Figure 16c, d).

When examining samples on a case by case basis, we observed that some dysplastic samples had a low copy number burden but highly rearranged genomes, notably dominated by translocations. The top three circos plots in Figure 17 show examples of this. The fourth case demonstrates a more representative level of rearrangement: a HGD patient with 70 SVs and 11.1% total CNA (proportion of whole genome affected by gain or loss using same criteria as in section 3.2.2). The fifth is a ND-NP for comparison with only 9 SVs and no CNAs.

Chromothripsis, chromosomal shattering due to mitotic segregation errors, is observed in 30% of OAC (Secrier et al., 2016) but this complex phenomenon has not been looked at in the pre-cancer stages apart from in a very small cohort of five HGD cases among which chromothripsis was observed in one case (Newell et al., 2019). In our cohort, chromothripsis was observed in one ND-NP case, but then from HGD onwards: 1/27 (3.7%) ND-NP; 0/12 (0%) ND-PP; 0/15 (0%) LGD; 4/26 (15.4%) HGD and 4/20 (20%) IMC. This confirms that these alterations are seen in the pre-malignant setting.

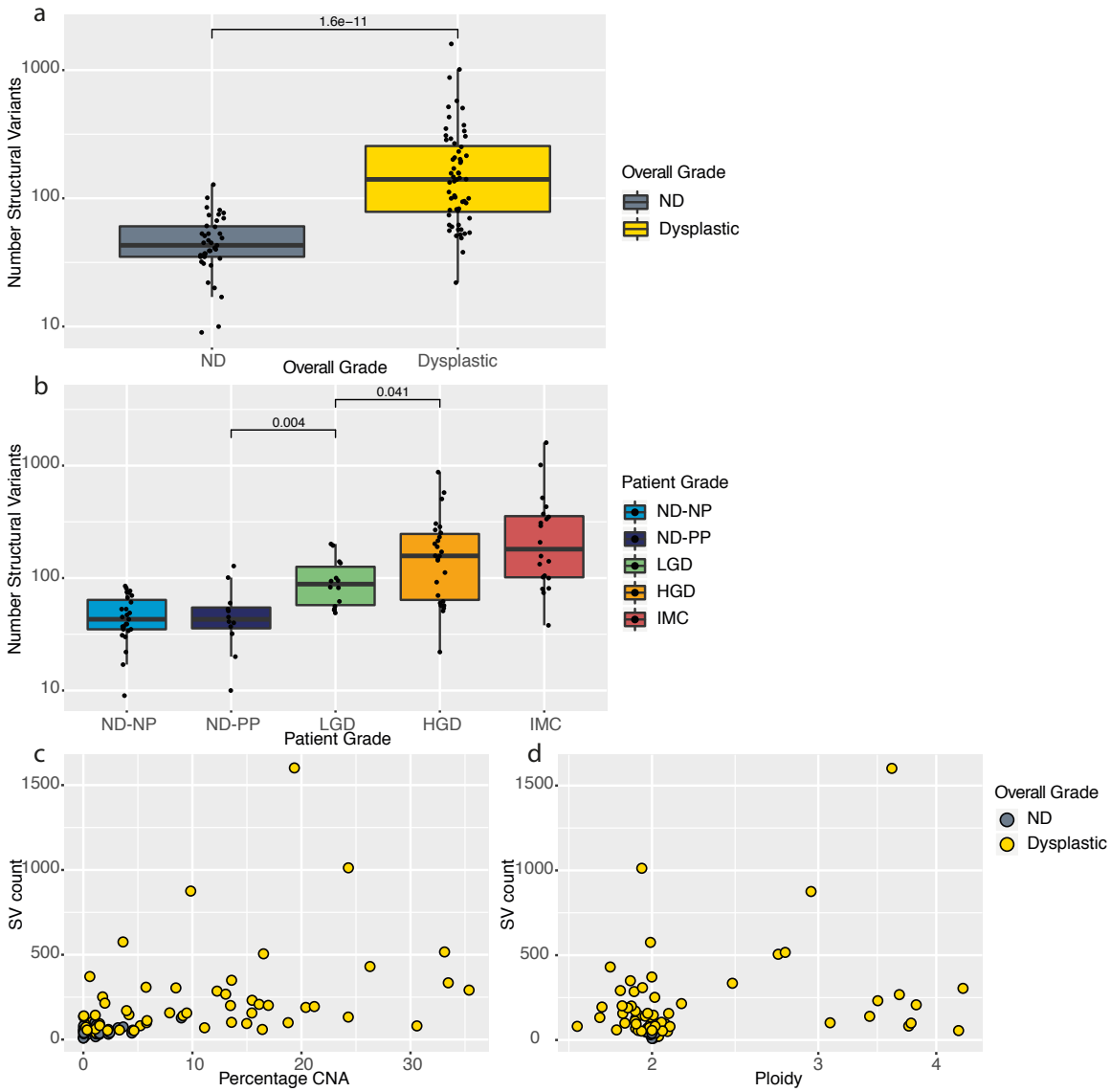


Figure 16 Structural variation across the grades

a. Total number of structural variants (SVs) (translocations, inversions, tandem duplications and large-scale deletions) grouped by overall grade. **b.** Total number of SVs by patient subgrade. **c. d.** Correlation between number of SVs and copy number aberrations (CNA) and ploidy per sample, coloured by dysplasia status. ND = Non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma. P value calculated using Wilcoxon Rank Sum test.

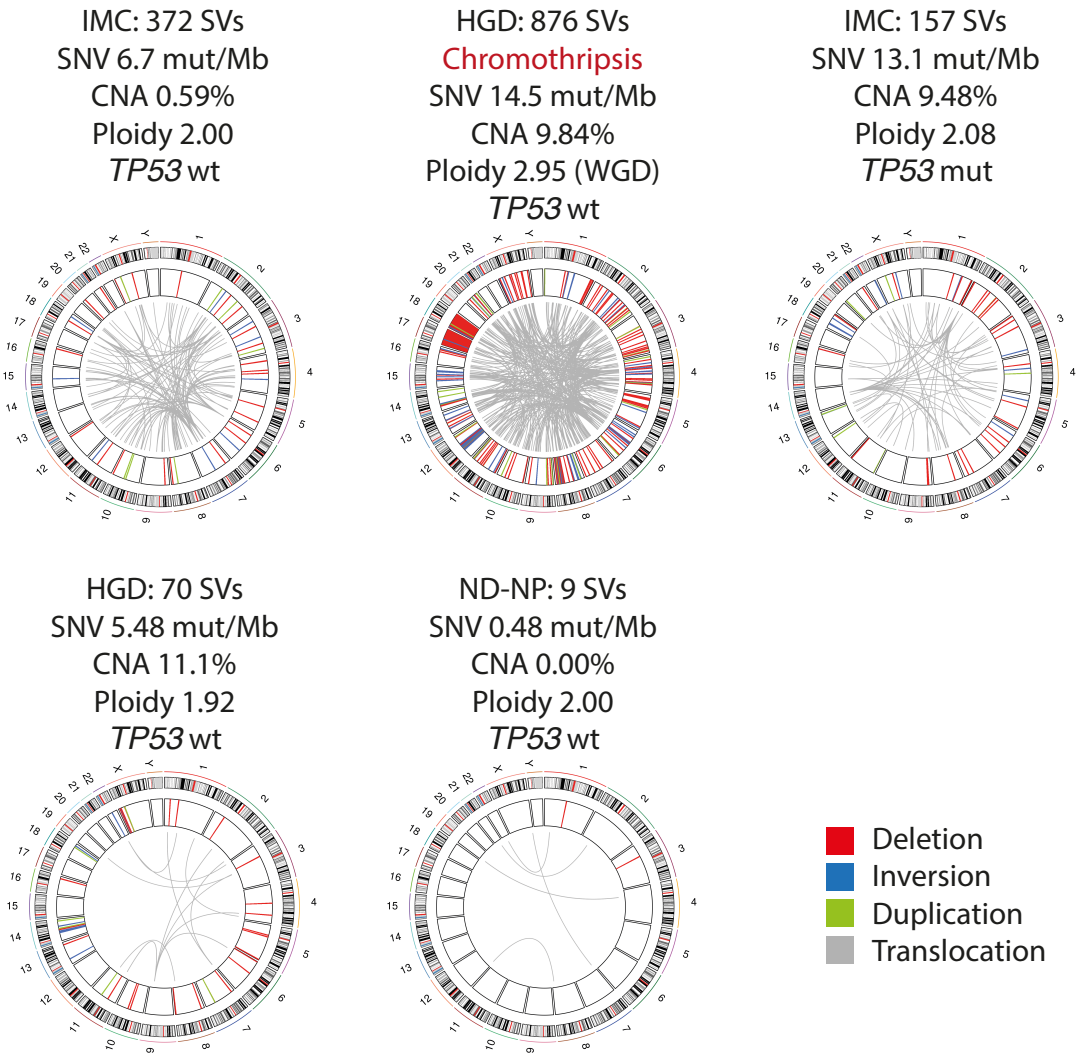


Figure 17 Circos plots of individual cases

Each outer circle represents the chromosomal ideogram. The inner wider circle displays the position and type of structural variant (red = deletion, blue = inversion, green = duplication). The middle circle tracks the translocation of one region to another part of the genome (grey). Grade of patient, mutation burden (SNV), percentage of copy number aberrations (CNA), ploidy and *TP53* mutation status are listed above each plot. The second case shows patterns consistent with chromothripsis. ND = Non-dysplastic, NP = non-progressor, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

Given the wide variation in the burden of SVs that we observed, we ordered the samples by this metric and found a gradual continuum with progression of dysplasia grade, with no apparent point at which a step-change occurred (Figure 18).

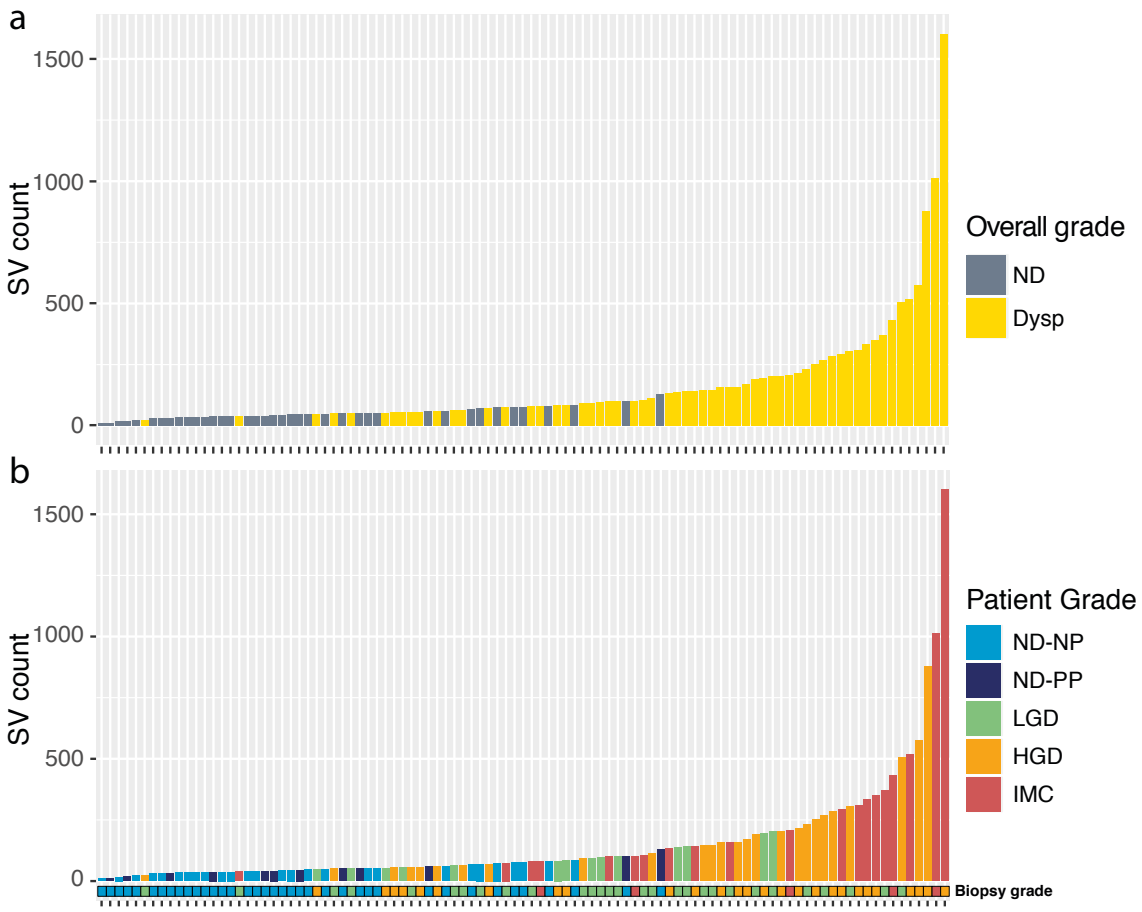


Figure 18 Samples ordered by burden of structural variants

Structural variant (SV) count plotted for each sample in order of total SV number. **a.** Samples colour coded by their overall grade. **b.** Samples colour coded by the patient grade with biopsy grade detailed below. ND = Non-dysplastic, Dysp = dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

The signatures of rearrangements, described initially by Nik-Zainal et al., can be considered in order to understand the patterns of SVs dominating the samples and to complement the base substitution signatures (Nik-Zainal et al., 2016). SVs can be classified as inside or outside of clusters. Clustered SVs occur when multiple breakpoints occur close together in the same region. SVs can then be classified by their type and size. The same non-negative matrix factorisation methods that are used for base substitution signatures can then be applied. We extracted five dominant signatures from our samples (Figure 19a). Rearrangement signature 1 is characterised by un-clustered translocations and deletions. Signature 2 is mainly un-clustered tandem duplications 10kb-1Mb. The SVs in signature 3 are clustered and dominated by very large (>1Mb) inversions, but also clustered deletions and tandem duplications. Signature 4 is comprised of un-clustered deletions and 5 is almost uniquely clustered translocations.

Signature 5 (clustered translocations) was seen across all grades (Figure 19b). Proportionally, it was higher in the samples with fewer SVs. However, the total number of SVs per case attributed to signature 5 was quite consistent throughout the grades: median counts (IQR) ND-NP 23 (16-31); ND-PP 16 (5-23); LGD 35 (26-46), HGD 41 (21-55), IMC 56 (22-82). Clustered translocations are likely to correspond to the fragile site regions of *FHIT* and *WWOX*, which we know are affected by CNAs in the ND stage. This suggests that these regions become focally affected by SVs early, with little further change with progression. Signature 2 (unclustered tandem duplications), in contrast, whilst also seen across the grades, is dominant in samples with low SV burdens. Signature 3 is dominated by samples with a high total number of SVs. Therefore, it appears that as samples become more rearranged, clustered deletions, tandem duplications and inversions are the predominant alterations that occur. Total numbers of these types of SVs may therefore be more useful than translocations, or an overall SV count, to classify the grade. There was no observable correlation between SV signature and *TP53* mutation status (Figure 19b: samples with *TP53* mutation marked with asterisks).

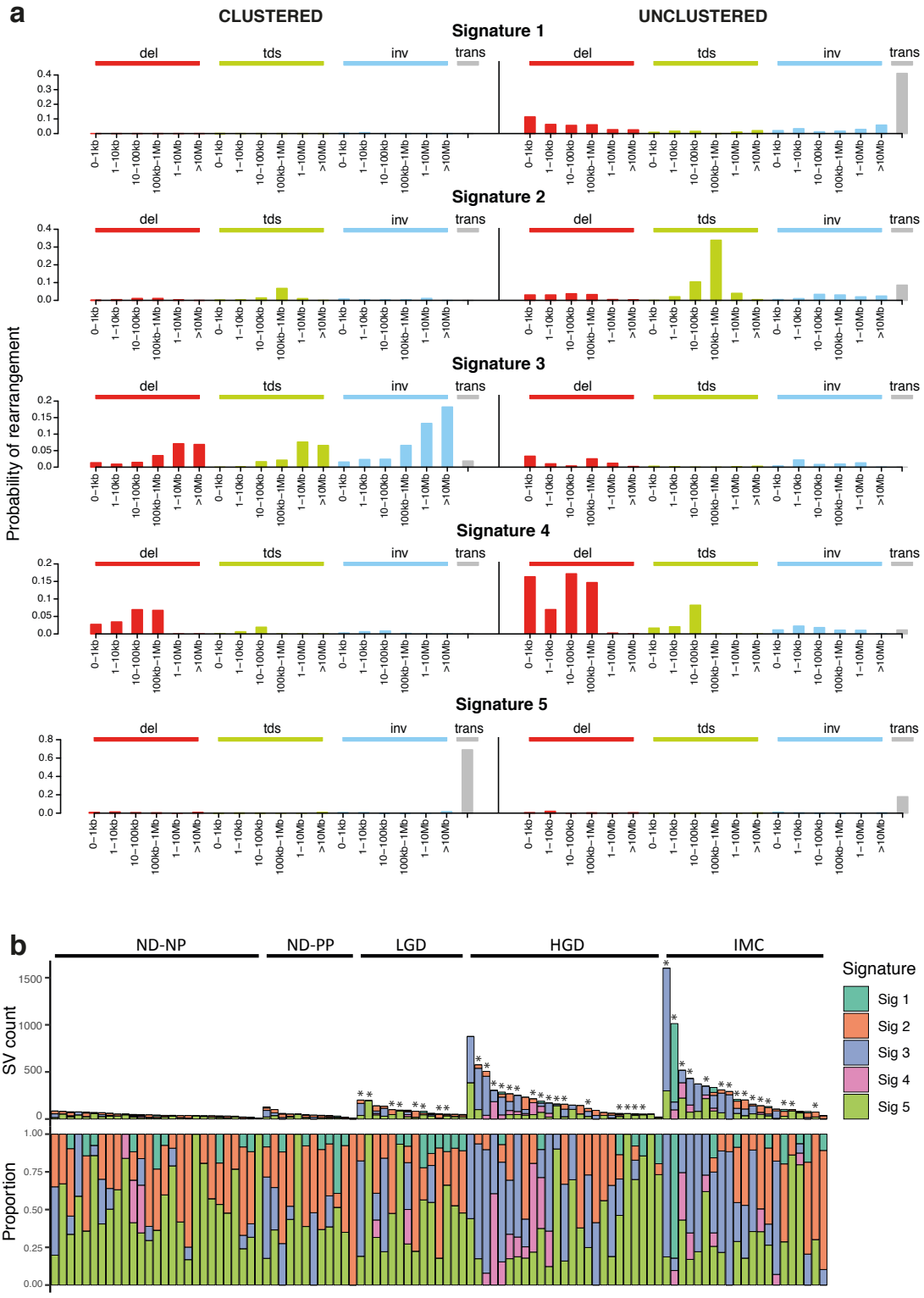


Figure 19 Structural variant signatures

a. The proportions and sizes of each structural variant (SV) type present in the signature. Split into clustered (left) and unclustered (right). Probability of rearrangement on y-axis. Rearrangement size on the x-axis. Del = deletion, tds = tandem duplication, inv = inversion, trans = translocation. **b.** The proportions of each SV signature contributing to our samples relative to total number of SVs. Samples grouped by grade and ordered by total number of SVs from high to low. Asterisks mark *TP53* mutant samples. ND = Non-dysplastic, Dysp = dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

3.3 Driver gene analysis in the progression of Barrett's oesophagus

In order to consider the specific genes that might be driving the progression of BE, a de novo analysis was performed to look for CN and SNV driver genes using GISTIC (Beroukhim et al., 2007), MutSigCV (Lawrence et al., 2013) and dNdScv (Martincorena et al., 2017). In addition to this, we considered the frequency of alterations in driver genes known to be altered in >10% OAC from our consortium (Frankell et al., 2019).

3.3.1 Copy number driver genes

For copy number drivers, we took the significantly amplified and deleted loci annotated by GISTIC and considered 1 million base pairs in each direction, in order to not miss any potential drivers in the flanking up and downstream regions. We then compared the expression of genes in wild type samples for a given gene versus those with a copy number change at that locus. Genes that were deleted in at least 5% of samples were included in our driver gene analysis. Eight of these genes were found to have a significantly lower expression between the deleted versus wild type samples (q value < 0.05). However, these eight genes all fell within the same GISTIC locus 9p21.3, containing *CDKN2A* (Figure 20, Figure 21). *CDKN2A* encodes p16INK4A which inhibits cyclin dependent kinases 4 and 6, thereby activating Rb protein and preventing G1-S phase cell cycle traversal. It is already known to be deleted early in BE and is a driver in OAC. *C9orf53* encodes an anti-sense RNA to *CDKN2A*. It was felt that the other genes in the region were passengers. No other deleted regions resulted in a significant loss of expression.

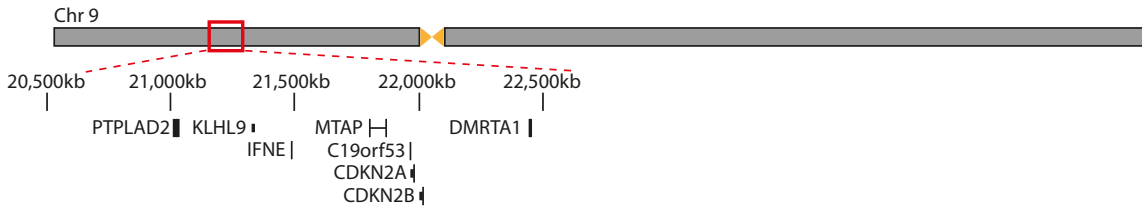


Figure 20 Genes with significantly reduced expression on 9p21.3

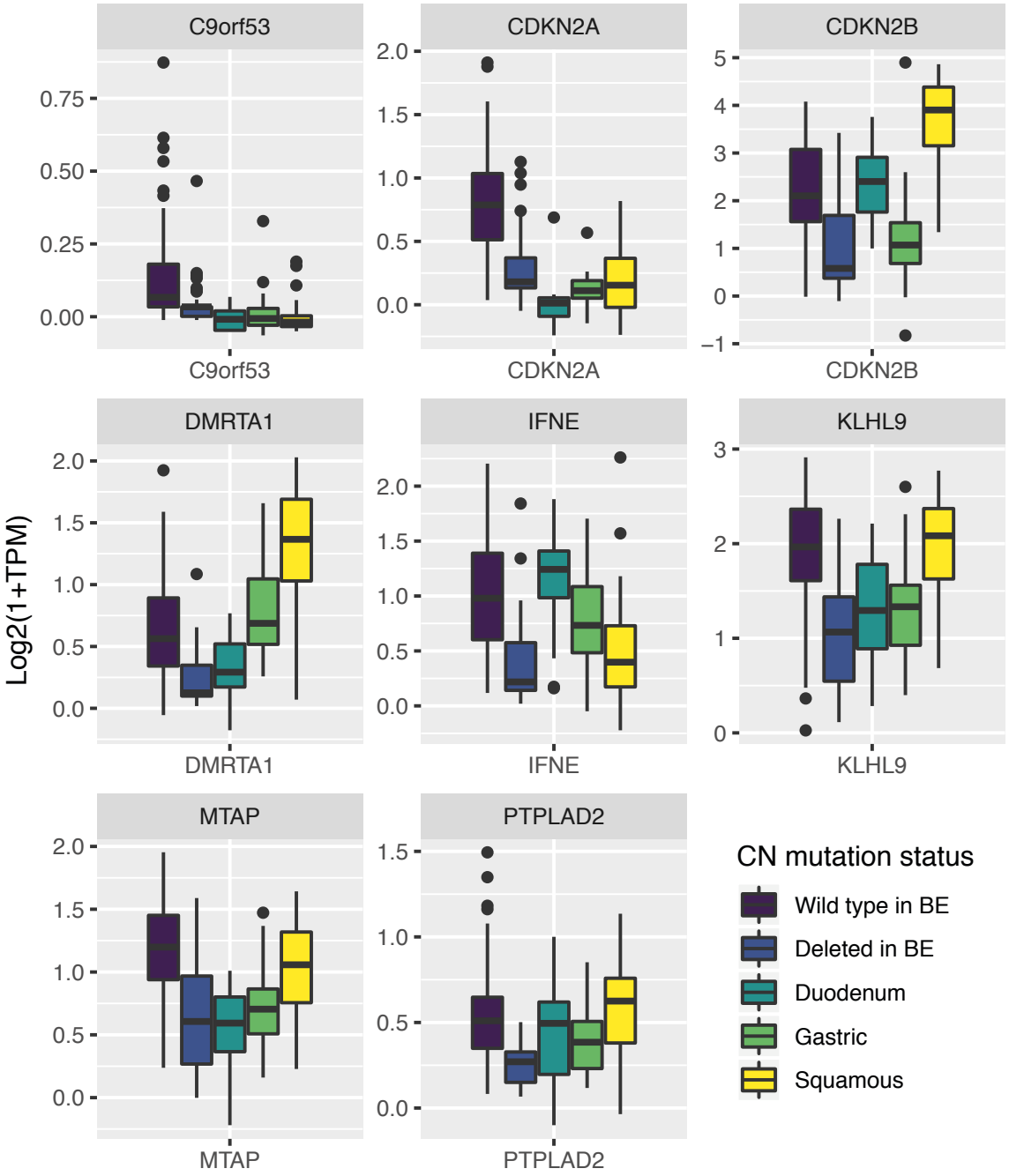


Figure 21 Expression of genes in significantly deleted regions

Eight genes with a significant difference (q value <0.05, False Discovery Rate; Wilcoxon Rank Sum for p value) in expression between deleted and wild type cases, where at least 5% of cases have a deletion. Log 2 expression plotted for each group. Comparison to expression in normal tissues. BE = Barrett's oesophagus, TPM = Transcripts Per Kilobase Million.

In contrast, using the same parameters for the amplified regions 57 genes were significant. To narrow it down we applied more stringent filtering to remove those with very low expression in the mutant samples (log2 expression <1). This reduced the list to 43 genes within 3 GISTIC loci 17q21.2, 17q12 and 18q11.2. However, a number of keratin genes were identified within these loci. These are genes expressed in squamous and so not likely to be relevant in malignant progression. They seemed to be masking any potential drivers but further stringent filtering e.g. q value < 0.05, did not help to remove them. It appeared that they were being called because we had initially extended the GISTIC regions to include 1 million base pairs up and down stream. Reducing the region to the original GISTIC loci successfully removed the majority of the keratin genes and reduced the gene list down to 17 genes within the three loci (Table 7).

Gene ID	Locus	No. of WT cases	No. of amplified cases	Mean expr in WT	Mean expr in amplified	Q value	Mean log2 fold change
TCAP	17q12	116	9	0.25	1.28	1.03E-03	5.11
STARD3	17q12	116	9	1.00	2.72	1.16E-04	2.72
PGAP3	17q12	116	9	1.33	3.35	1.10E-04	2.51
GRB7	17q12	115	10	1.22	2.86	9.69E-05	2.35
MIEN1	17q12	116	9	1.23	2.73	4.79E-03	2.21
CDK12	17q12	118	7	1.87	3.77	5.52E-04	2.02
ERBB2	17q12	116	9	2.94	5.32	1.16E-04	1.81
PPP1R1B	17q12	116	9	2.82	5.04	3.43E-04	1.78
KRT17	17q21.2	113	11	1.06	2.83	9.13E-04	2.68
LEPREL4	17q21.2	114	10	1.17	2.67	7.80E-05	2.27
RARA	17q21.2	118	7	1.24	2.47	1.20E-03	2.00
EIF1	17q21.2	114	10	3.34	5.71	7.80E-05	1.71
KRT15	17q21.2	113	11	1.19	2.02	6.63E-03	1.70
JUP	17q21.2	113	11	4.47	7.20	3.30E-05	1.61
KRT19	17q21.2	113	11	7.78	9.54	4.64E-04	1.23
MIB1	18q11.2	115	10	2.58	3.33	6.24E-03	1.29
GATA6	18q11.2	113	12	3.02	3.88	2.50E-02	1.28

Table 7 Genes with a significant difference in expression in amplified cases versus wild type cases (q value<0.05).

Ordered by log2 fold change in amplified versus wild type (WT). q value<0.05 considered significant. (Wilcoxon Rank Sum test for p value). Expression calculated as log2(1+TPM). TPM = Transcripts Per Kilobase Million.

Loci 17q12 and 18q11.2 contained *ERBB2* and *GATA6* respectively, both known amplified drivers in OAC. We looked at the functions of the other genes in these 2 loci. The only two with apparently interesting functions were *GRB7* and *CDK12*. *GRB7* encodes a growth factor receptor which interacts with EGFR and promotes the activation of downstream MAP kinases STAT3, MAPK1 and MAPK3. *CDK12* regulates transcriptional elongation and genes involved in DNA repair. Thereby, it is required for the maintenance of genomic stability. Its downregulation activates the MAPK pathway (Iorns et al., 2009).

Locus 17q21.2 did not contain a known driver. Two genes had potentially relevant functions: *RARA* encodes the retinoic acid receptor which is a transcriptional factor with roles in cell differentiation and proliferation, especially known for its role in acute promyelocytic leukaemia (PLM-RAR translocation). *JUP* is a paralogue of *CTNNB1* and is a junctional plaque protein in desmosomes and intermediate junctions (http://www.uniprot.org/uniprot/P14923#function).

We plotted the expression of amplified versus wild type cases against normal tissues duodenum, stomach and squamous oesophagus for each of these genes (Figure 22). The elevated expression of *JUP* and *GRB7* in squamous did not make them promising drivers.

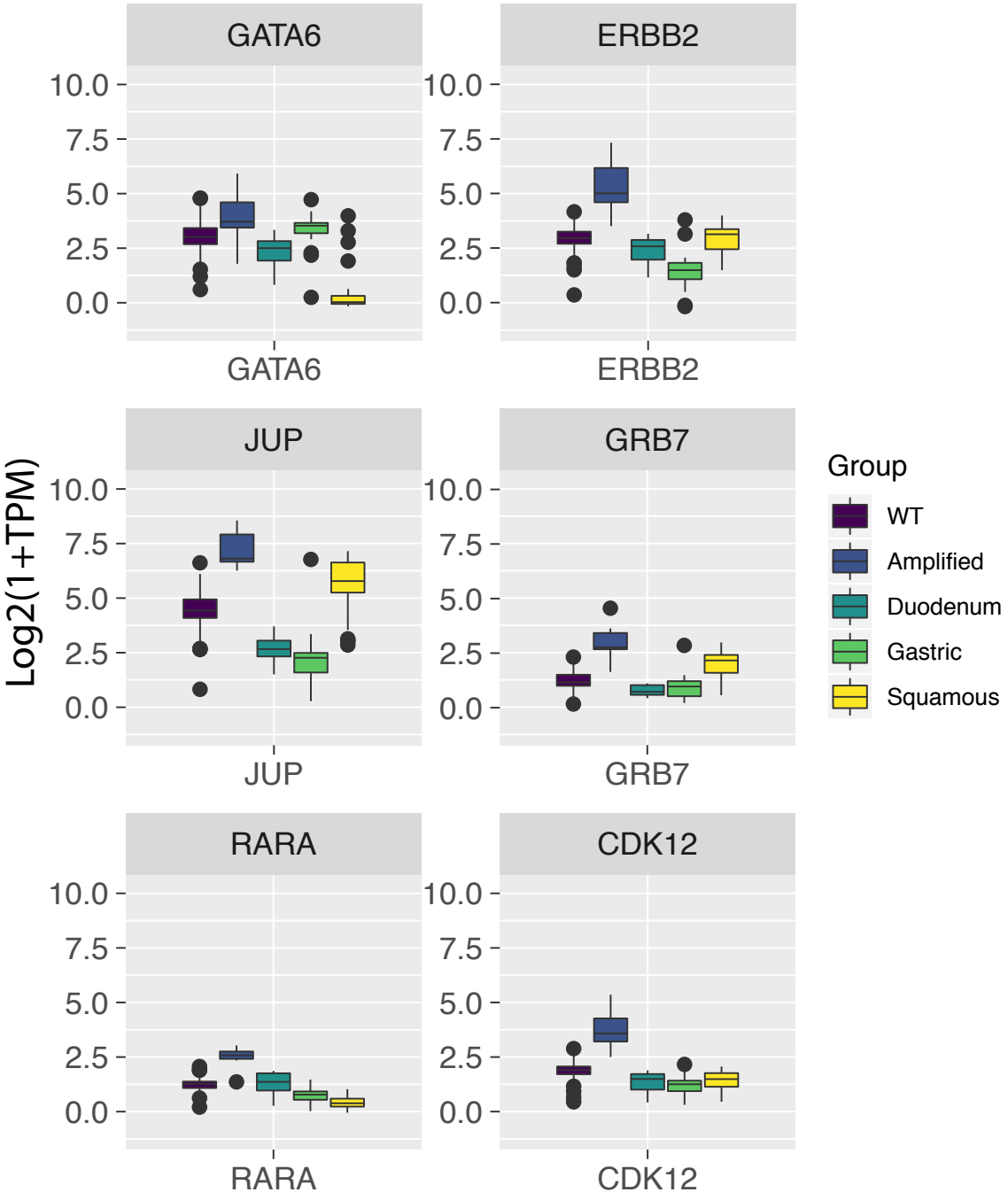


Figure 22 Expression of 6 significantly amplified genes in driver gene discovery

Six genes with cancer-related functions and a significant difference in expression between amplified and wild type (WT) cases (q value <0.05, False Discovery Rate; Wilcoxon Rank Sum for p value), where at least 5% of cases have an amplification and log 2 expression > 1 in amplified samples. Log 2 expression plotted for each group. Comparison to expression in normal tissues. TPM = Transcripts Per Kilobase Million.

Next we examined the correlation between expression and copy number as shown in Figure 23. The log2 expression for each sample is plotted against the log2 of CN corrected for ploidy. A sample with 2 copies and a ploidy of 2 will have a log2 (CN/ploidy) of 0. Each dot represents one sample. It is known that there is often only a weak correlation between amplification and expression because of the potential closed state of the chromatin. *GRB7*, *RARA* and *CDK12* exhibited a wide variation in expression with normal copy number and ploidy and this did not specifically help to identify them as drivers. Two samples had a log2 expression > 4 and more than 16 copies for *ERBB2*. We examined this in more detail and found it to be driven by SVs, with a concentration of duplications within the region of chromosome 17, a proportion of which were likely to be extra-chromosomal as double minutes (Figure 24).

Overall, none of the four potential new drivers (*JUP, GRB7, RARA, CDK12*) feature in the COSMIC Cancer Gene Census (https://cancer.sanger.ac.uk/census) nor have been found to be drivers in 551 OAC cases (Frankell et al., 2019). We also showed that they were all co-amplified with *ERBB2* and not amplified independently in other samples. Given this evidence that they were unlikely to be drivers, and instead passengers to *ERBB2* amplification, we decided to proceed using the copy-number drivers already reported as amplified in OAC (Frankell et al., 2019).

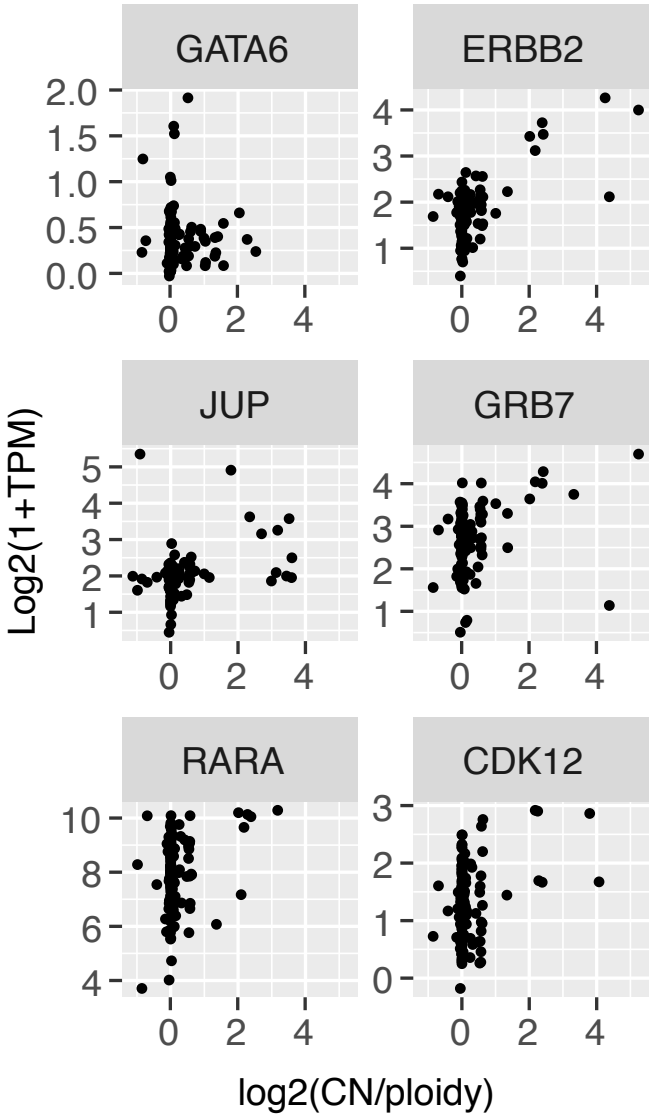


Figure 23 Expression versus copy number of 6 significantly amplified genes in driver gene discovery

Log2 expression for each sample plotted against the total copy number (CN) divided by ploidy. i.e. a sample with 2 copies and a ploidy of 2 will have a log2 (CN/ploidy) of 0. Six genes plotted: each with cancer-related functions and a significant difference in expression between amplified and wild type (WT) cases (q value <0.05, False Discovery Rate; Wilcoxon Rank Sum for p value), where at least 5% of cases have an amplification and log 2 expression > 1 in amplified samples. TPM = Transcripts Per Kilobase Million.

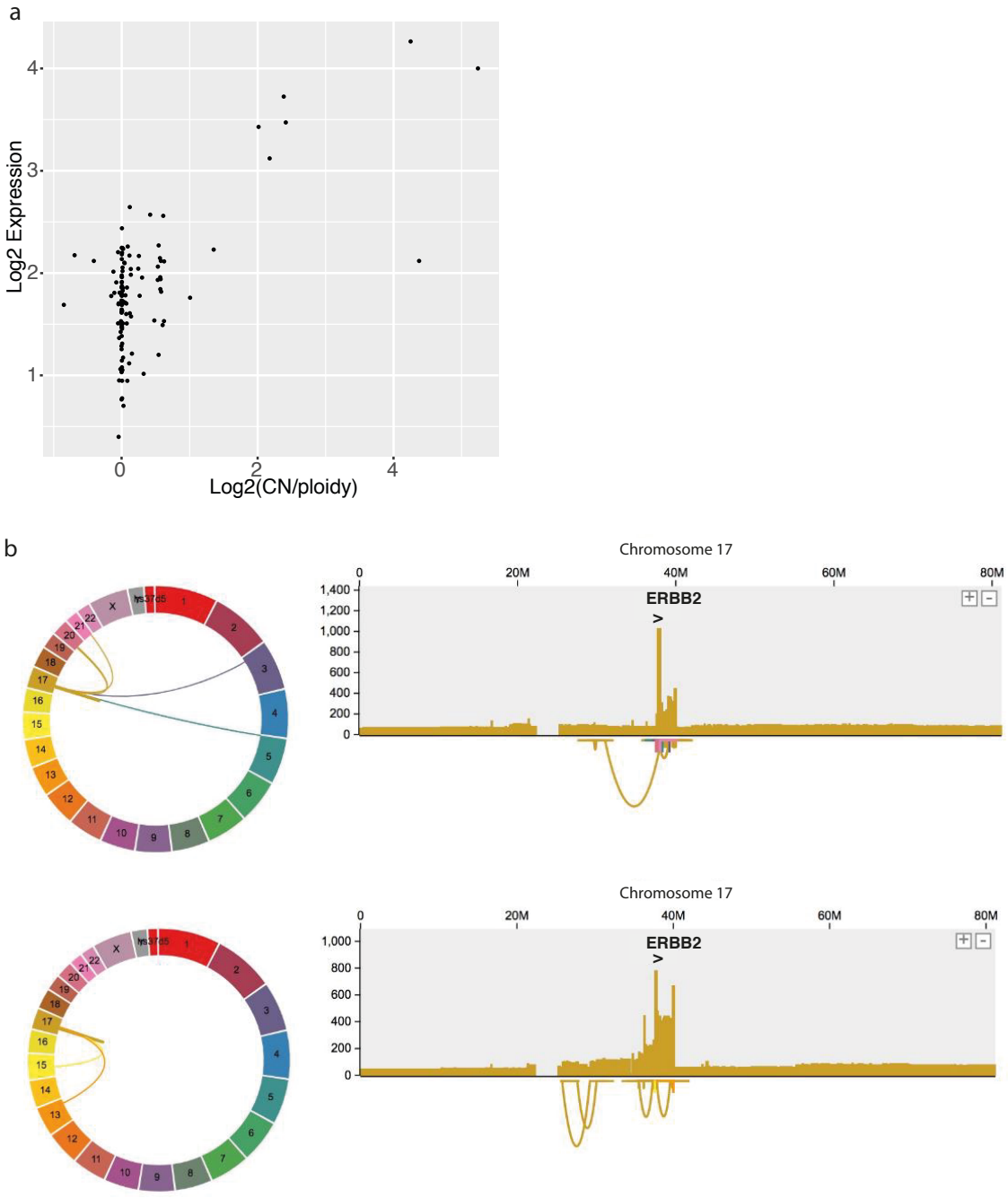


Figure 24 SV-driven ERBB2 amplification

a. Scatterplot showing correlation of *ERBB2* copy number and expression levels in all samples. **b.** Two high grade dysplasia cases with >16 copies of *ERBB2* due to duplication structural variants. Circos plots on the left show positions of translocations from the *ERBB2* locus. Only the SVs related to *ERBB2* have been plotted.

3.3.2 Point mutated driver genes

We used MutSigCV and dNdScv to identify potential point mutations in driver genes in the cohort, i.e. mutations in genes which lead to clonal expansion and are positively selected for in growth of the lesion. MutSigCV identifies genes mutated more often than expected by chance. It uses the mutational process of neighbouring genes with similar genomic properties to model the background mutation rate. Whereas dNdScv uses evolutionary methods to detect genes under positive selection in cancer by comparing the ratios of non-synonymous to synonymous mutations within genes. dNdScv identified *TP53, CDKN2A, ARID1A* and *PAIP2B* as drivers. MutSigCV additionally identified *SNTG2, SMARCA4, PTEN* and *MUC6*. These genes are all known drivers in oesophageal cancer (Frankell et al., 2019). We did not expect to identify any new drivers because very large cohorts would be needed to identify low frequency driver events. But, furthermore, if a gene is not driver in OAC, it is unlikely to be functionally important in BE progression.

We took these 8 genes from our discovery but also added any driver genes mutated in at least 10% of OACs, discovered by Frankell *et al.*, 2019. We compared the frequency of point mutations/indels, amplifications and deletions of these genes across the grades in our cohort and observed how the driver landscape changed with progression (Figure 25).

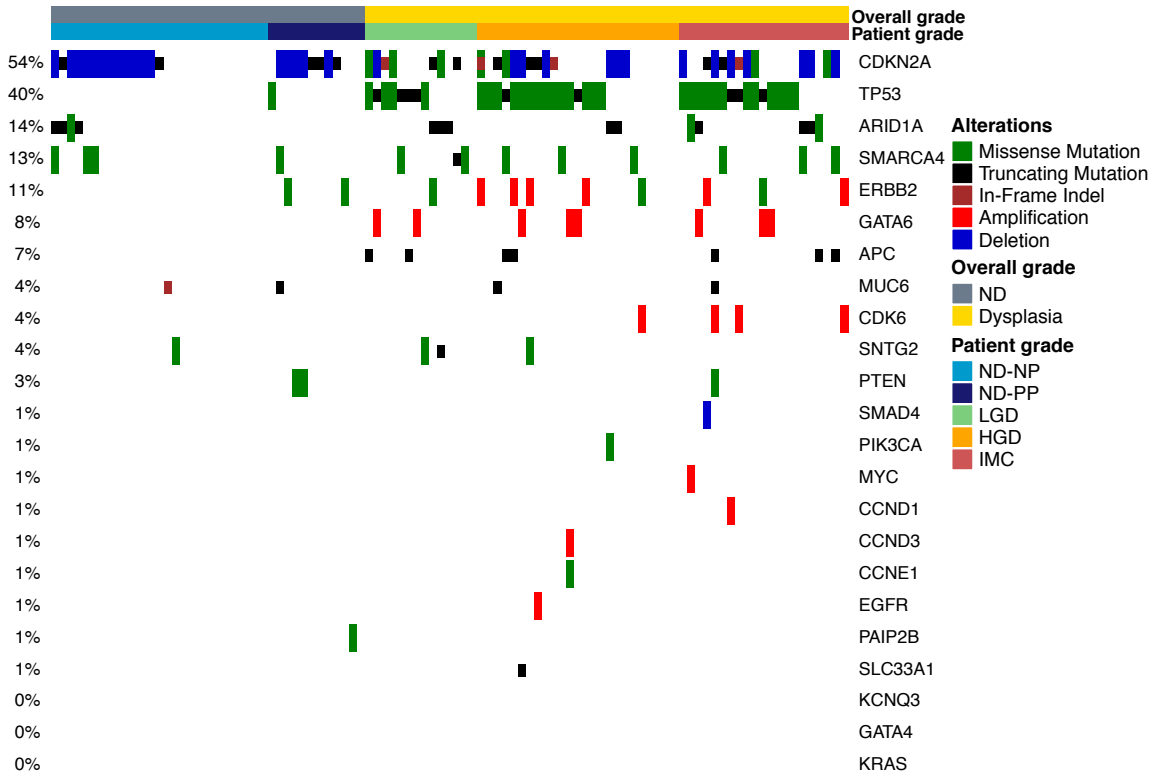


Figure 25 Driver gene mutation frequency in the cohort

Each sample is represented vertically, ordered by grade. Genes are ordered by frequency and colour coded for type of alteration. ND = Non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

CDKN2A was the most frequently altered gene, occurring in 54% of the cohort. This was predominantly by deletion, as found in the CN driver analysis. It was altered in a similar proportion of both ND and dysplastic cases (56.4% ND vs. 51.7% dysplastic). It has been shown to be mutated in only 29% of OAC (Frankell et al., 2019). This raises the question as to why there is this drop and whether clones with a *CDKN2A* alteration are perhaps less likely to expand. However, this has not been investigated here.

TP53 was mutated in 65.0% (39/60) dysplastic cases and 0% non-progressor cases (only point mutations were considered and not LOH). However, a mutation was seen in one pre-progressor case (ND). This case was examined in further detail and although all biopsies were graded by the histopathologist as ND in 2009, this patient had had biopsies graded as indefinite in 2008 and 2006, and LGD in 2005, prior to routine p53 staining. The H&E slides were recalled for this case and were re-graded by two senior Consultant pathologists and p53 staining was performed. The 2005 LGD was downgraded to ND, however, there was significant nuclear p53 staining in the 2006 and 2008 biopsies. The 2006 biopsy was upgraded to LGD and the 2008 biopsy to HGD. The abnormal areas must have been missed on the endoscopy in 2009: highlighting the risk of sampling bias in surveillance. This patient was, therefore, considered as HGD and excluded from the group. It can be clearly seen that *TP53* mutation is coincident with dysplasia (0% ND-NP, 0% ND-PP, 57.4%, LGD, 64.0% HGD, 71.4% IMC) (Figure 26). This is in keeping with previous work from our lab which compared ND and HGD samples with a gene panel and showed *TP53* to be present in HGD (72%) and OAC (69%) but not NDBE (Weaver et al., 2014). This was the only driver gene for which the frequency of mutation increased stepwise with grade. In OAC wild type *TP53* cases are often affected by an *MDM2* mutation, however we did not observe this here.

Other genes which were altered at a lower frequency were also confined to the dysplastic stages, just not in a stepwise manner. For some drivers we saw mutual exclusivity between genes within the same pathways: *ARID1A* and *SMARCA4* are both members of the SWI-SNF family and involved in chromatin remodelling. Both were mutated in all grades, at similar rates to seen in cancer (Frankell et al., 2019), although *ARID1A* mainly by truncating mutations and *SMARCA4* by missense mutations. *ARID1A* mutation was also mainly mutually exclusive to *TP53* mutation, as has been previously described in OAC and other cancers e.g. (Frankell et al., 2019; Guan et al., 2011).

GATA6 and *ERBB2* amplification were also mutually exclusive and confined to dysplasia. However, missense mutation of *ERBB2* was also seen in 16.7% ND-PP (2/12) but not ND-

NP. Several genes were mutated in similar proportions as seen in OAC: *ERBB2* alteration is seen in OAC (16%); *GATA6* was altered in similar numbers of LGD, HGD and IMC, and has been shown to be mutated in 14% OAC. *APC* and *CDK6* were also only altered in dysplasia. *APC* was altered in 11.7% (7/60) of dysplasia samples (9% in OAC) and *CDK6* in 7% (4/60) (14% in OAC).

KRAS and *MYC* have both been shown to be the joint third recurrently mutated genes in 19% of OAC (Frankell et al., 2019). We did not see any *KRAS* mutations in our cohort. *MYC* was mutated in one IMC case. These genes may be important in progression to an invasive phenotype.

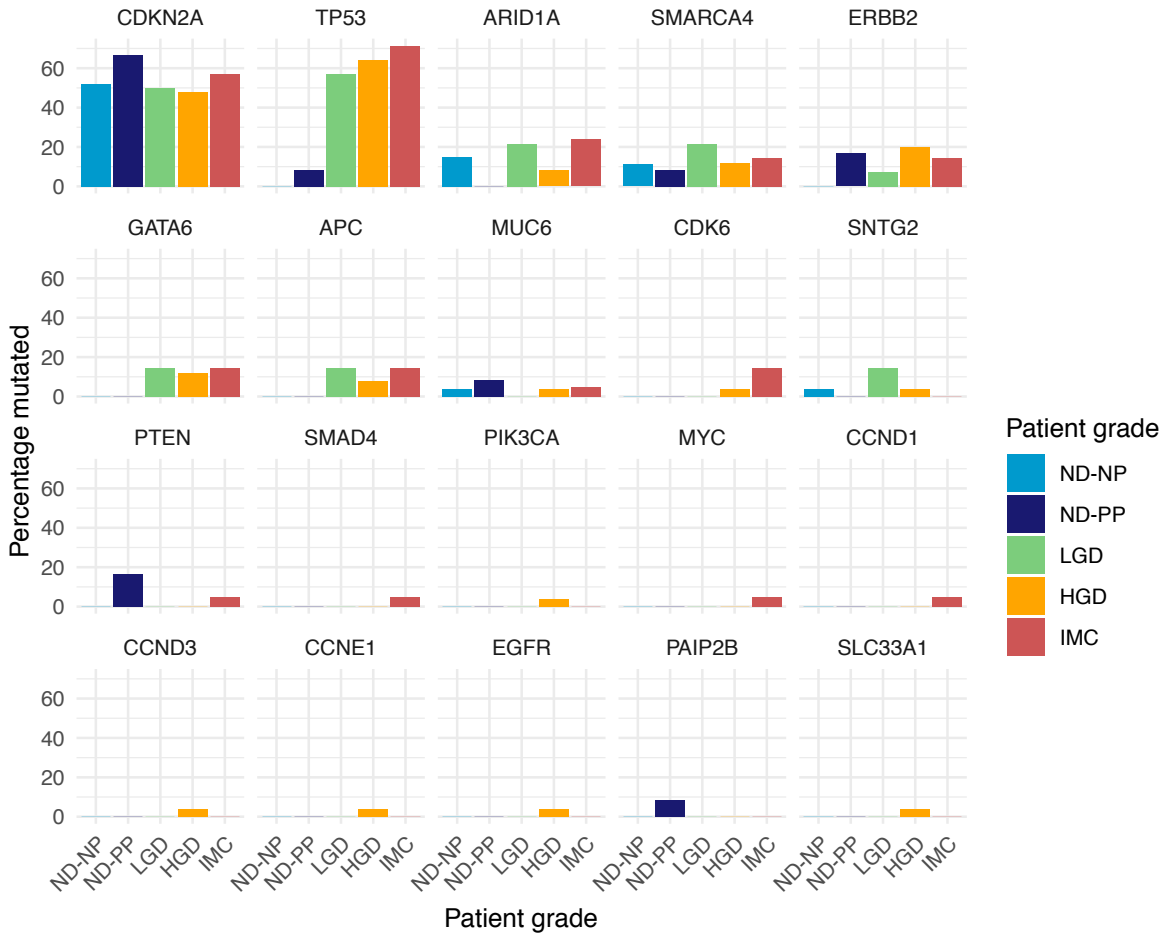


Figure 26 Frequency of driver gene alteration per grade

Proportion of samples per grade with an alteration in each of the identified driver genes. ND = non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

To summarise the analysis up to this point, we considered all of the above features for each individual case across the cohort. Figure 27 shows the genomic and clinical features for each case. They are ordered by total number of SVs, detailing CNA, SNVs and numbers of drivers. We ordered by SV number because they had the widest variance across the cohort. The heatmap gives clinical and demographic information, with the first row (overall grade) grouping the samples into ND (grey) and dysplastic (yellow). This reiterates the findings above in which SVs lie on a continuum in the ND to dysplastic progression. There is an overall trend towards an increase in CNAs, SNVs and driver events with increasing dysplasia. However, an individual sample may be dominated by one genomic feature which tips the balance to progression to dysplasia e.g. low SV count but high % total CNA. This highlights the heterogeneity of progression, with many potential parallel pathways to cancer.

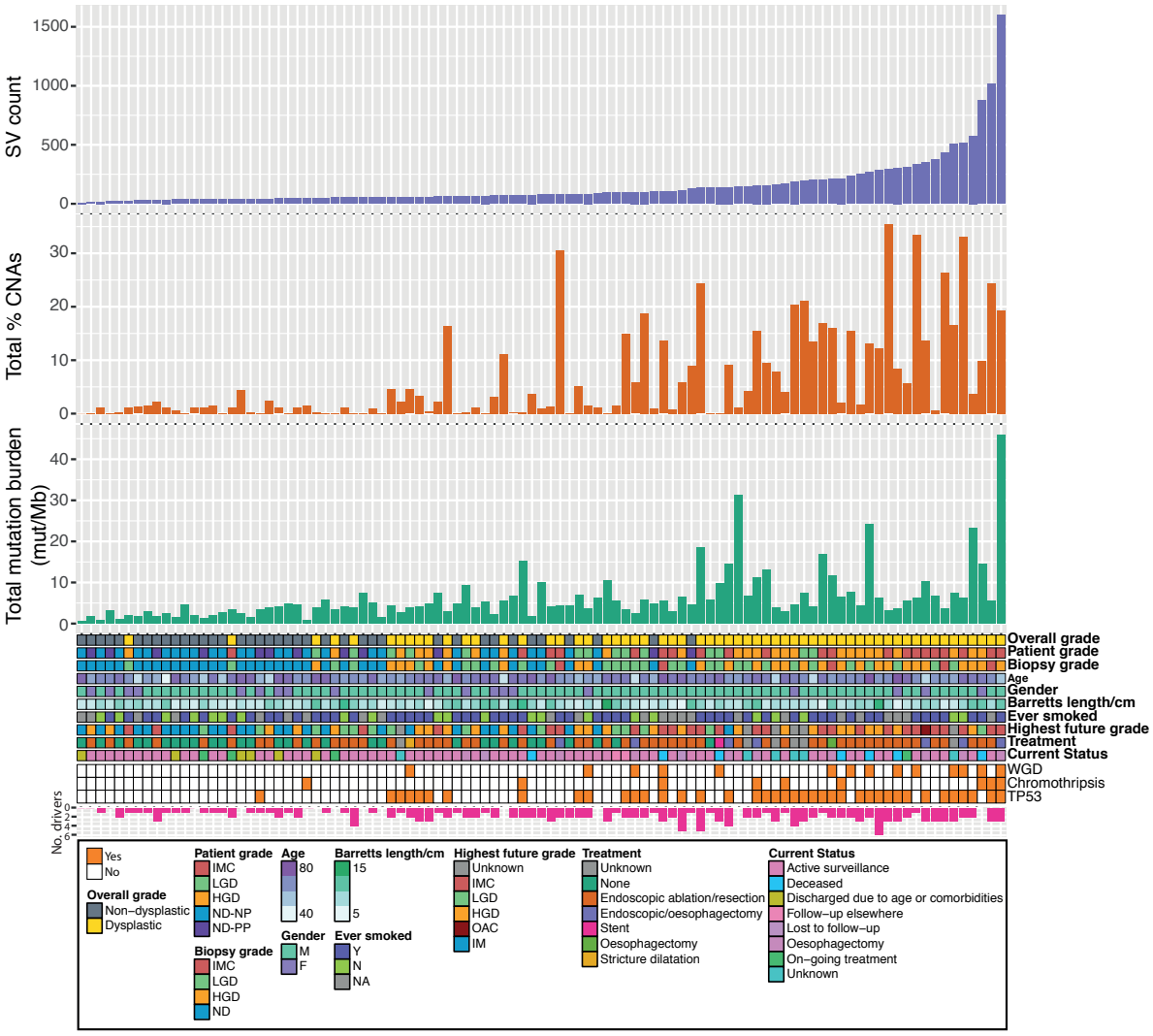


Figure 27 Summary of genomic and clinical features of the pre-cancer cohort

All samples ordered by total number of structural variants (SV). Total copy number aberrations and mutation burden plotted. The heatmap details clinical features and other genomic features of each sample. The number of driver genes per sample plotted in pink. WGD = whole genome duplication, ND = non-dysplastic, NP = non-progressor, PP = pre-progressor, IM = intestinal metaplasia, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma, OAC = oesophageal adenocarcinoma.

3.3.3 Barrett's oesophagus adjacent to cancer (Trio BE)

The majority of previous research using whole genome or exome sequencing has been conducted on non-dysplastic BE sampled from adjacent to tumour, usually from the oesophagectomy resection specimen (Ross-Innes et al., 2015a; Stachler et al., 2015). These samples represent the final stage in cancer evolution, in which large clonal expansions are likely. Hence it is not known to what extent the adjacent BE samples recapitulate the earlier time points in the natural history of the disease. Therefore, samples taken from BE adjacent to cancer (Trio BE) were compared with the pre-cancer BE cohort. Since the adjacent samples had various grades of dysplasia (38 ND, 1 indefinite, 5 LGD, 3 HGD) the ND samples were compared to the ND pre-cancer BE, and the dysplastic Trio BE to the dysplastic pre-cancer group.

Figure 28 shows the genomic profiles comparing the pre-cancer cohort to the Trio BE. There was no significant difference between the median mutational burden, CNA or SV count when comparing the two ND groups or the two dysplastic groups (Figure 28a-c). However, for SVs, it was noticed that the range was wider in the Trio ND BE than the pre-cancer ND (Figure 28d) with some samples having SV counts similar to dysplastic samples: Trio ND BE SV range 2-323 (median 57); pre-cancer ND range 9-128 (median 43). Three Trio BE cases (6.4%) had WGD: two of which were HGD (one of which was *TP53* mutant) and one ND (*TP53* wild type).

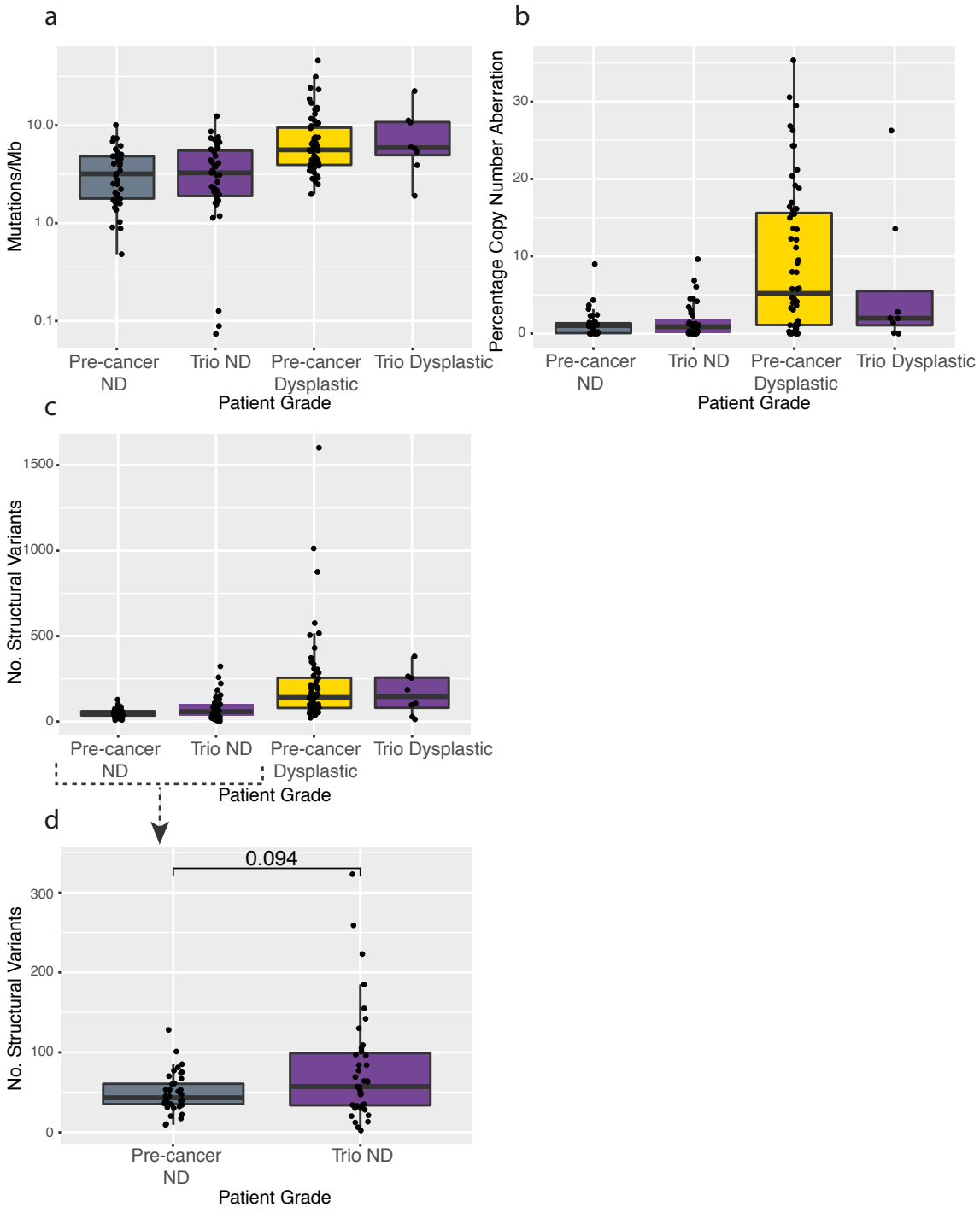


Figure 28 The mutational landscape of Barrett's adjacent to cancer compared to non-adjacent

Box plots showing **a.** mutation burden, **b.** copy number aberrations (CNA) and **c.** structural variant (SV) burden in pre-cancer Barrett's oesophagus (BE) and BE adjacent to cancer (Trio) split into non-dysplastic (ND) and dysplastic. **d.** SV counts for pre-cancer ND and Trio ND only. P value calculated by Wilcoxon Rank Sum test. ND = non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia.

In the pre-cancer cohort, mutational signatures had been preserved across the grades, with signatures 17 a and b positively correlating, and signature 1 negatively correlating, with mutational burden. If the mutational signatures form early in the natural history of BE, and do not change over the course of progression, then we would expect to see a similar signature profile in the Trio BE adjacent to cancer. Taking only the ND Trio BE (n=39), we compared the mutational signatures observed with the patterns in the ND-NP and PP (Figure 29a). There was no difference in the mutational signature proportions in the ND Trio BE compared to the other ND BE. This also suggests that the tumour does not influence the mutational signature profile of the surrounding BE.

The SV signatures in the ND Trio BE had a higher proportion of Signature 3 (clustered inversions, deletions and tandem duplications) in the samples with more SVs (Figure 29b). This was not seen in the pre-cancer ND samples, as they had lower SV counts. So, the ND Trio BE was more similar to the dysplastic pre-cancer samples, despite their benign phenotypic pathological appearance.

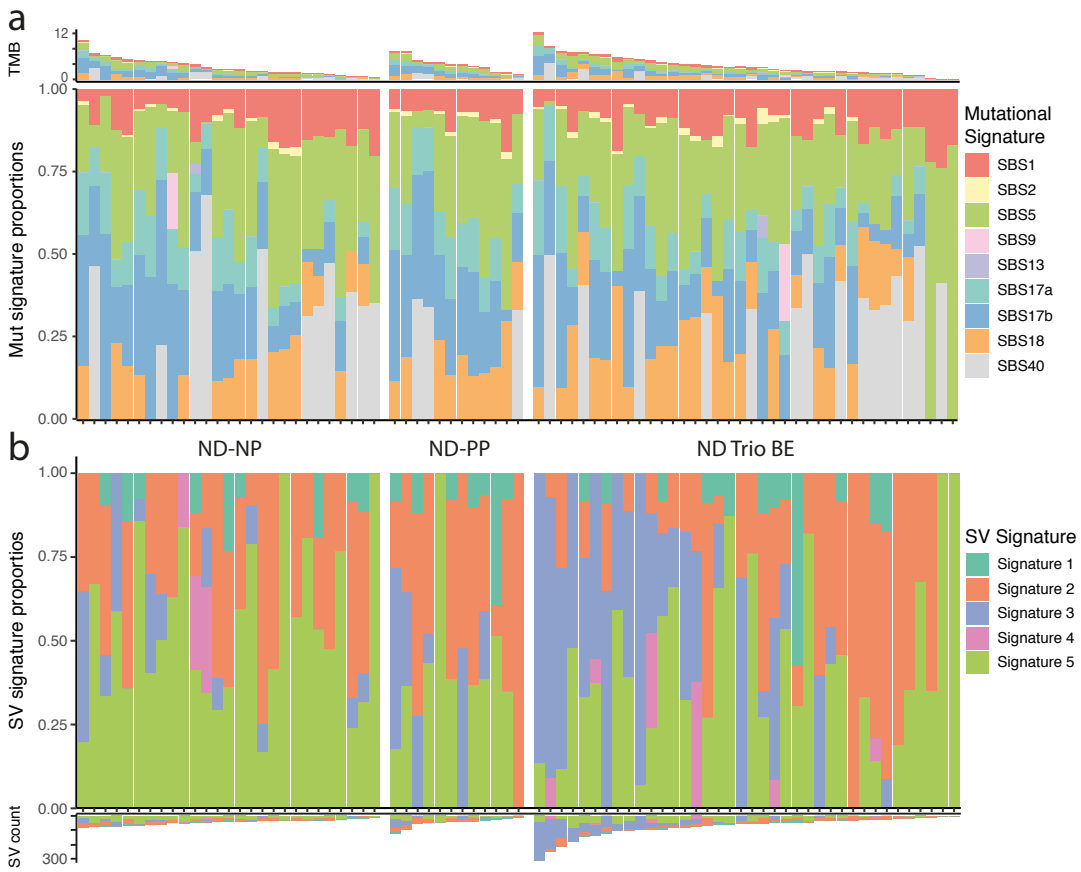


Figure 29 Mutational and SV signature proportions in non-dysplastic BE adjacent to cancer compared to pre-cancer non-dysplastic BE

a. Each mutational signature plotted as a proportion of all SNVs per sample. Total mutation burden (TMB) plotted above. Only pathologically non-dysplastic (ND) samples included. **b.** Structural variant (SV) signatures plotted as a proportion of total number of SVs for each case. Total SV number per case plotted below. NP = non-progressor, PP = pre-progressor, SBS = single base substitution.

In the pre-cancer cohort, we had observed the SVs to lie on a continuum across dysplasia grades. We therefore examined where the Trio BE cases would lie on this continuum. We found them to span the pre-cancer cases, irrespective of grade, falling along the whole continuum. Figure 30 shows, again, that the ND Trio BE can be more affected by structural variants than pre-cancer ND samples, and appear more genomically similar to dysplastic tissue. Chromothripsis was observed in two Trio BE, both ND and both wild type for *TP53*. The circos plots for these are shown below in Figure 30. The figure also highlights one ND Trio with 323 SVs, dominated by translocations.

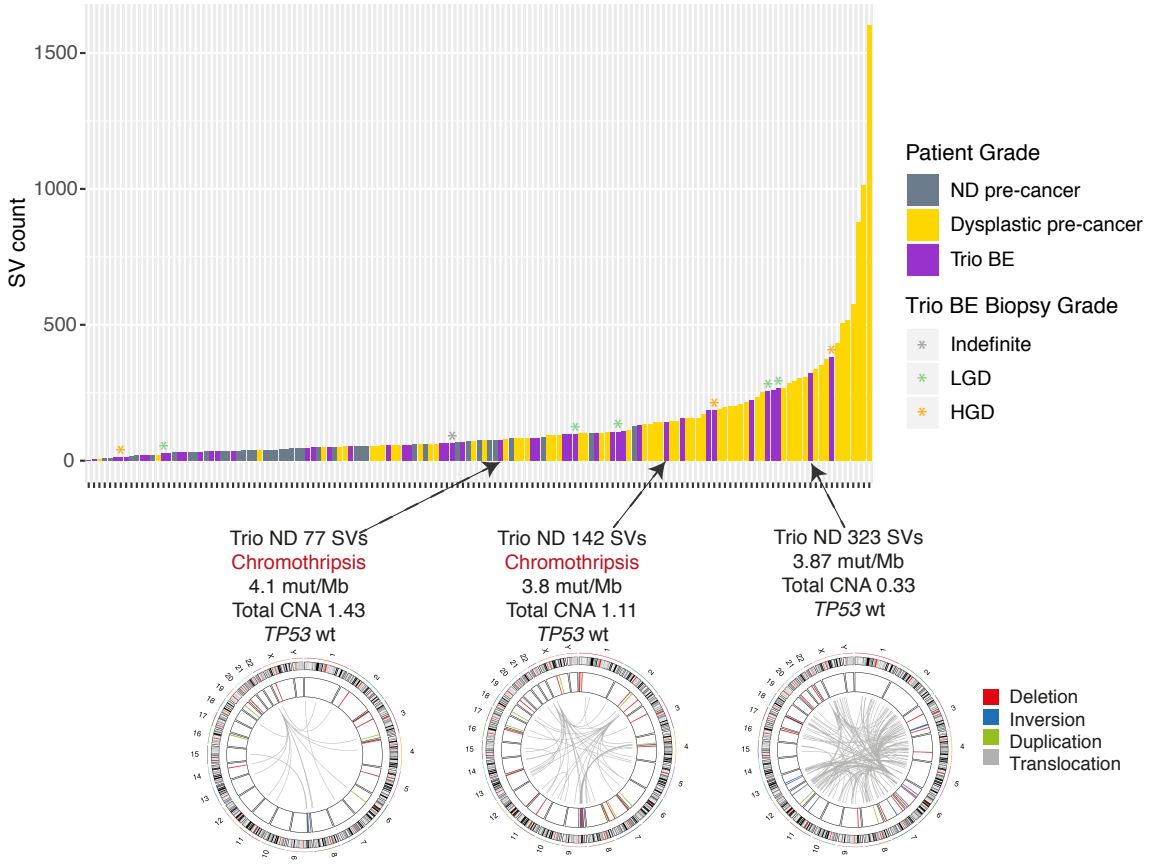


Figure 30 Structural variation continuum with Barrett's adjacent to cancer (Trio) plotted

Structural variant (SV) count plotted for each sample (x axis). Samples colour coded by their overall grade. Dysplastic Trio BE marked with a colour-coded asterisk. Circos plots shown to highlight the rearrangements in three ND Trio BE samples: two exhibiting chromothripsis and one with more than 300 translocations. Chromosomes represented around the circumference of the plot. Deletions, duplications and inversions marked on the inner circle, with translocations as grey lines. ND = non-dysplastic. BE = Barrett's oesophagus.

Given the wide range of SV counts seen in the ND Trio BE, we speculated whether there might be a difference in the mutational profile of Trio BE cases with a similar mutation spectrum to its adjacent cancer and those without. We compared the mutational overlap between BE/tumour pairs and considered there to be overlap if >= 10% of the total number of BE mutations were shared (Figure 31). BE biopsies with mutational overlap with their adjacent tumour were not more likely to be dysplastic (p value = 0.55; Fisher's Exact test). There was no significant difference between the mutational burden or SV count of BE with mutational overlap with adjacent tumour and BE without (Figure 32). There was a trend towards a significant difference in total CNAs although the median number CNAs were low: Trio BE with >10% shared mutations: range 0.003-26.7% (median 1.95%.); Trios with <10% shared mutations range 0-6.0% (median 0.69%), (p value = 0.018 Wilcoxon Rank Sum test). So overall, it did not appear that the more rearranged clones in adjacent BE were necessarily the ones from which the tumour had arisen.

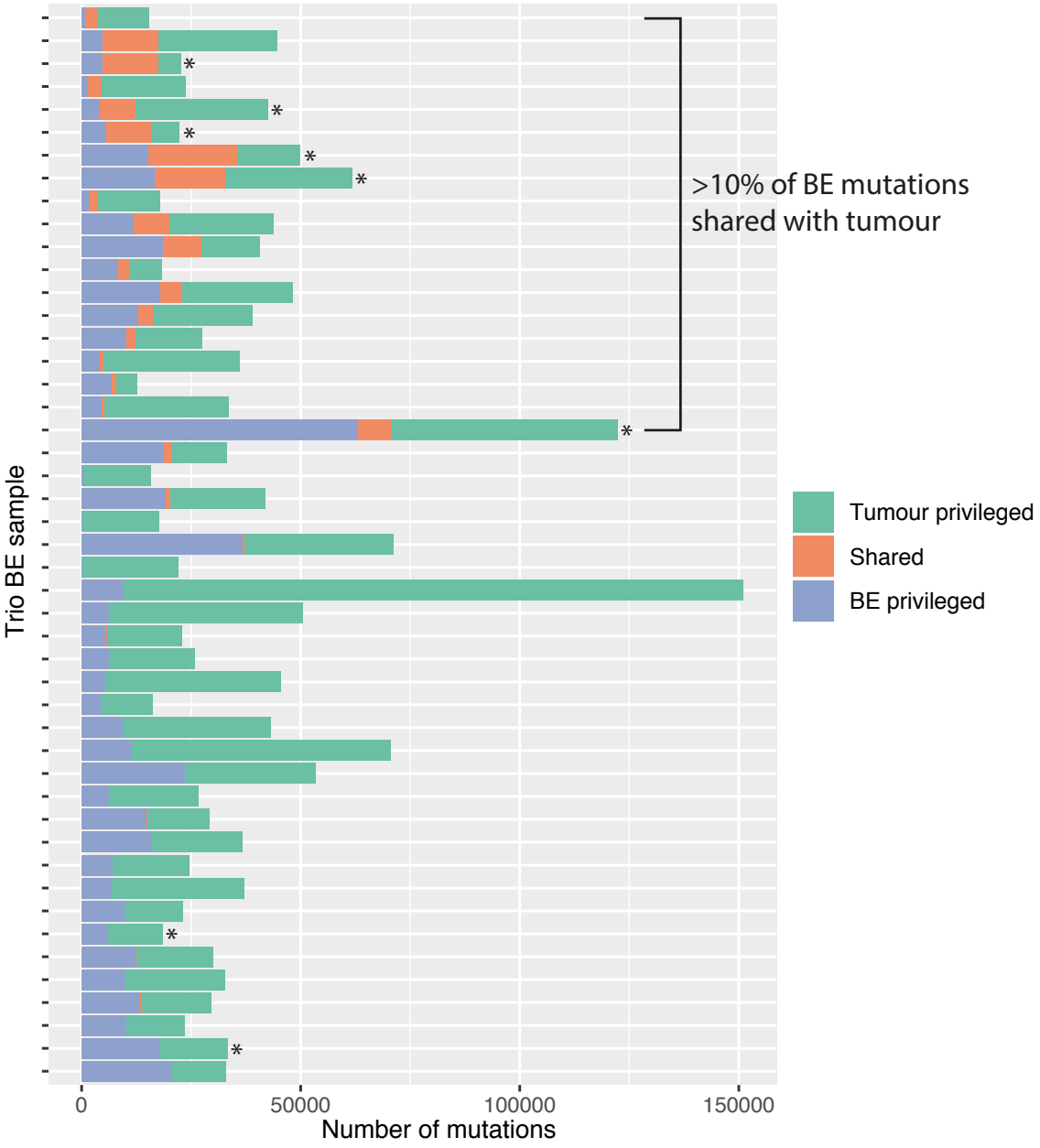


Figure 31 Mutational overlap between Trio Barrett's oesophagus and adjacent tumour

Bar plot showing total number of mutations called for each Trio Barrett's oesophagus (BE) sample and its adjacent cancer, and the number of mutations shared. BE samples marked with asterisks are dysplastic. Only BE sharing at least 10% of its mutations with the adjacent cancer was considered as an overlap.

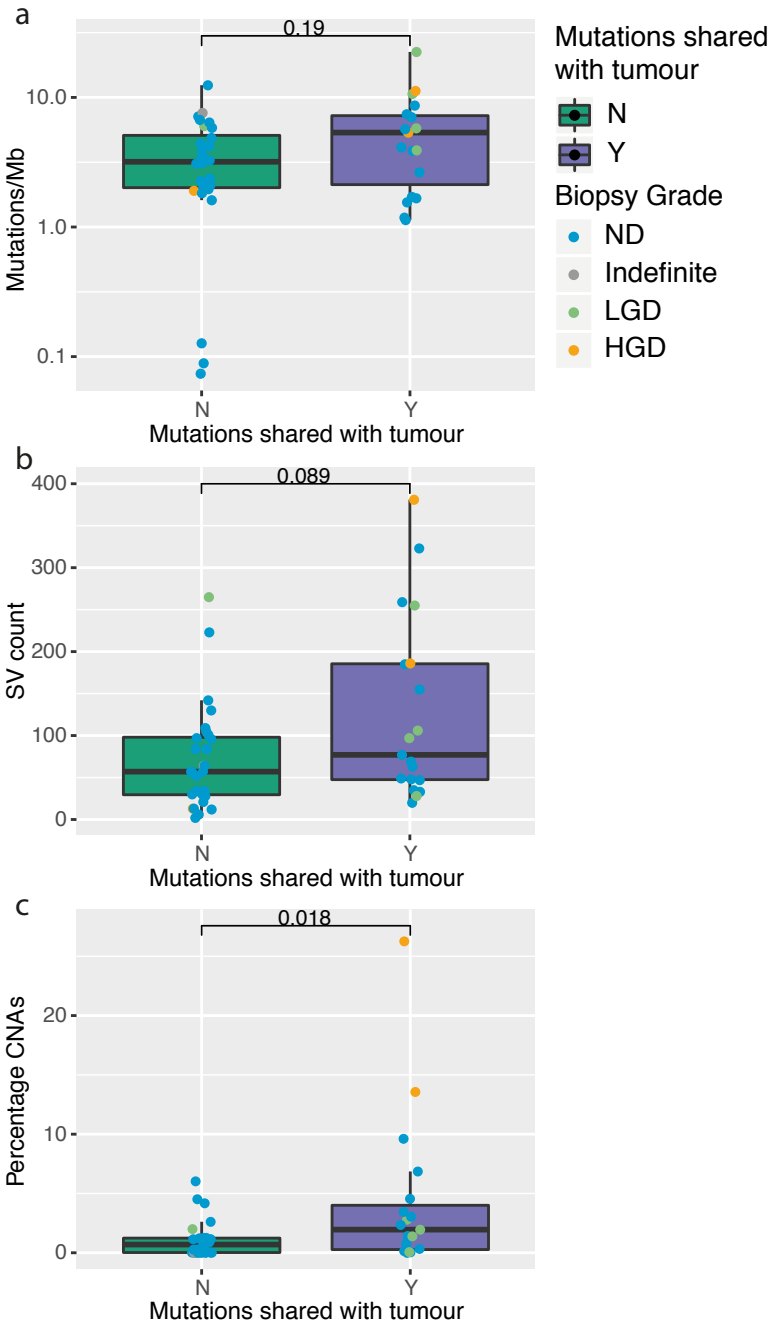


Figure 32 Genomic profiles of Trio BE sharing mutations with adjacent tumour

a. Mutation burden, **b.** structural variant (SV) count and **c.** percentage of copy number aberrations (CNA) plotted for Trio Barrett's oesophagus sharing >10% mutations with its adjacent tumour and those not. P values calculated by Wilcoxon Rank Sum test. ND = non-dysplastic, LGD = low grade dysplasia, HGD = high grade dysplasia.

Lastly, we considered whether the driver gene alterations were occurring at different frequencies in the Trio ND BE compared to the NP and PP (Figure 33). Again, we considered only the Trio BE with ND histology (n=39) for a direct comparison and removed the *TP53* mutant ND-PP that on re-review had been from a patient with dysplasia. The *CDKN2A* alteration rate was similar to the pre-cancer ND BE, implying that it is an early event in progression. We found 15% (6) of the ND Trios to be *TP53* mutant. This was lower than the rate seen in the dysplastic pre-cancer BE but it was perhaps unexpected to see it in phenotypically-ND samples. We also observed mutations in other driver genes that were otherwise only mutated in IMC or cancer: *SMAD4, GATA4, KRAS*. So, whilst the Trio ND BE overall had a similar genomic profile to ND pre-cancer BE in terms of mutation burden, CNAs and SVs, the driver gene profile was more similar to early cancer. It has not previously been shown that non-dysplastic BE next to a tumour is not representative of true ND, non-progressing BE.

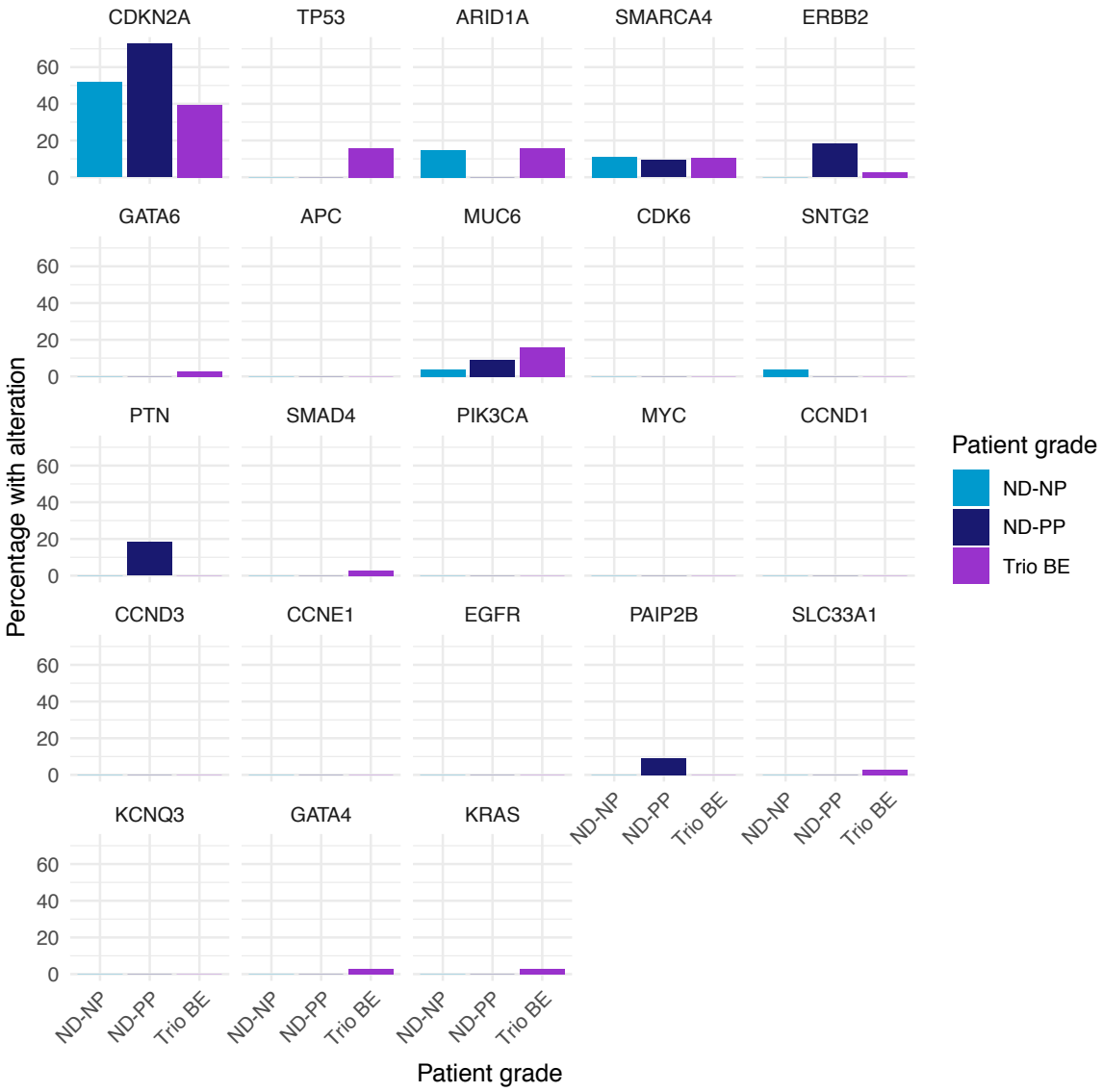


Figure 33 Driver gene alteration frequencies in only the non-dysplastic Barrett's adjacent to cancer

Proportions of driver genes with amplification, deletion or point mutations in non-dysplastic non-progressor samples, pre-progressor samples and from BE adjacent to cancer.

3.4 Summary

We created a carefully-curated cohort of BE samples, representing the grades from non-dysplastic through to IMC, for genomic sequencing. This was with the aim of elucidating the key biological processes driving Barrett's oesophagus (BE) to progress to oesophageal adenocarcinoma (OAC). This is the first whole genome sequencing focussing on the different grades of disease progression, and with long follow-up. Whilst, overall, we observed a significant increase in median mutation burden and percentage CNAs between ND and dysplastic samples, the ranges were wide. Analysis did not reveal a clear separation between either the overall grades of the patients or the actual grades of the biopsies sequenced. It has previously been shown that non-dysplastic BE is highly mutated, with similar mutation rates to OAC (Ross-Innes et al., 2015a) so it is perhaps unsurprising that there was little change in the mutation burden with progression through the grades. Total % CNAs, in contrast, was low for ND samples but had a wide variation for dysplastic cases: with some IMC samples displaying very few CNAs. We noticed that, in these cases, a high number of structural rearrangements were present instead, potentially explaining the phenotype seen. Total SV number per sample had the biggest variance within grades. Although there were no defining genomic features separating NP and PP samples, analysis of expression data may reveal differences.

Placing all the samples in order of their total SV count revealed a clear continuum from ND to dysplastic samples which was consistent with the morphological grade and clearer than using other genomic features. HGD and IMC samples had the highest burdens of SVs. Structural variants in BE have not been widely studied because of the need for whole genome sequencing and this progression with grade has not previously been observed. Signatures 5 (clustered translocations) and signature 2 (unclustered tandem duplications) were seen across all grades. Whereas signature 3 (clustered deletions, inversions and tandem duplications) was mainly observed in dysplastic samples with high SV burdens. In some cases, SVs were driving some of the CNAs that we were seeing e.g. *ERBB2* amplification. The next step will be to perform an in-depth analysis of regions recurrently affected by SVs which may reveal new, or higher frequencies of, driver genes in progression. It would also be very interesting to build a larger cohort of non-progressors and pre-progressors with multiple timepoints to track how the SV count changes over time and observe how early we see these large-scale alterations.

An analysis of driver events in the grades confirmed a stepwise increase in frequency of mutation of *TP53* with grade. We did not see *TP53* mutation in the ND-PP samples. This is in keeping with previous work from our lab (Weaver et al., 2014) but contrasts a recent study which used a gene panel on NP and PP and found a 46% *TP53* mutation rate in the pre-progressors (Stachler et al., 2018). *GATA6*, *ARID1A* and *APC*, although mutated at lower frequencies, were confined to dysplasia and not seen in ND biopsies. *ARID1A* mutation was mutually exclusive with *TP53* mutation. We did not see any differences between the long-term ND-NP cases and the ND cases that went on to progress. This suggests that the mutations and rearrangements that are accumulate with progression are not there at the early stages. However, we appreciate that only one biopsy was sequenced per case, so we do not necessarily get a complete insight.

The analysis of ND BE sampled from adjacent to cancer (Trio BE) had a similar mutational burden and proportion of copy number aberrations to non-progressor and pre-progressor BE. It has previously been assumed that ND tissue from adjacent to cancer would be genomically similar to ND-NP but this has previously neither been confirmed or refuted. Whilst the TMB and CNAs did not differ, a number of Trio BE samples had very rearranged genomes and a higher frequency of driver gene alterations, otherwise not seen at the ND stage. This highlights the likelihood of an initial field cancerisation, which has been widely described, or, furthermore, the possibility of the surrounding BE being altered by the presence of the tumour. The BE adjacent to cancer, therefore, should not be compared to pre-cancer non-progressing BE and, in these cases, the histology is misleading.

Overall, the findings confirm the heterogeneity of BE oesophagus. No specific, recurrent molecular features define the assigned phenotypic, pathological grades. Instead, BE appears to be more a continuum of disease. Plotting all the features of all samples per case suggests that there are multiple paths to cancer: one dysplastic sample may have a dominance of structural rearrangement and another may have more focal CNAs.

Given the mutation burden and clonal/origins of BE metaplasia, it is an interesting point of discussion as to whether the histology of metaplasia is misleading and if BE should be regarded as a neoplasia instead. However, the inability of most BE to progress/grow counteracts this and one could argue that the terminology fits the clinical trajectory.

This heterogeneity makes finding biomarkers to diagnose progression difficult. Using cut-offs of burdens of molecular features may prove to be more useful than panels of specific molecular features and this shall be explored further in Results 4.

4. Results 2: The transcriptomic landscape of Barrett's oesophagus

Aim 1: Elucidate the key biological processes driving Barrett's oesophagus (BE) to progress to oesophageal adenocarcinoma (OAC) by performing an integrated analysis of genomic and transcriptomic sequencing of the individual grades of BE.

- Analyse the differential expression of genes in non-dysplastic versus dysplastic Barrett's oesophagus in order to identify recurrent expression changes in progression.
- Perform pathway analyses for up- and downregulated genes to understand the key processes occurring in progression.
- Compare how the immune composition of the biopsy correlates with grade.

4.1 Sample comparison of the most variably expressed genes

Of the samples that had undergone WGS, 93 Barrett's oesophagus (BE) samples (25 non-dysplastic (ND) non-progressors (NP), 10 ND pre-progressors (PP), 12 low grade dysplasia (LGD), 25 high grade dysplasia (HGD) and 21 intramucosal carcinoma (IMC) had sufficient quantities and adequate qualities to undergo whole transcriptome sequencing. In addition, we sequenced 31 BE adjacent to cancer (Trio BE) and 80 normal tissues as a comparison (18 duodenum (D2), 38 normal oesophagus (NE) and 24 gastric cardia (GC)).

After performing batch correction using ComBat, group comparisons were made by calculating the 1000 most variably-expressed genes and performing a principal component analysis (PCA). Firstly, considering only the normal tissues, they formed distinct groups of samples which confirmed that the sequencing was of good quality (Figure 34a). One gastric sample clustered with squamous and vice versa. Pathology records did not suggest contamination and they had not been extracted on the same day or plated on the same batch. Secondly, PCA demonstrated that neither batch nor RNA integrity number (RIN) were confounding factors (Figure 34b, c).

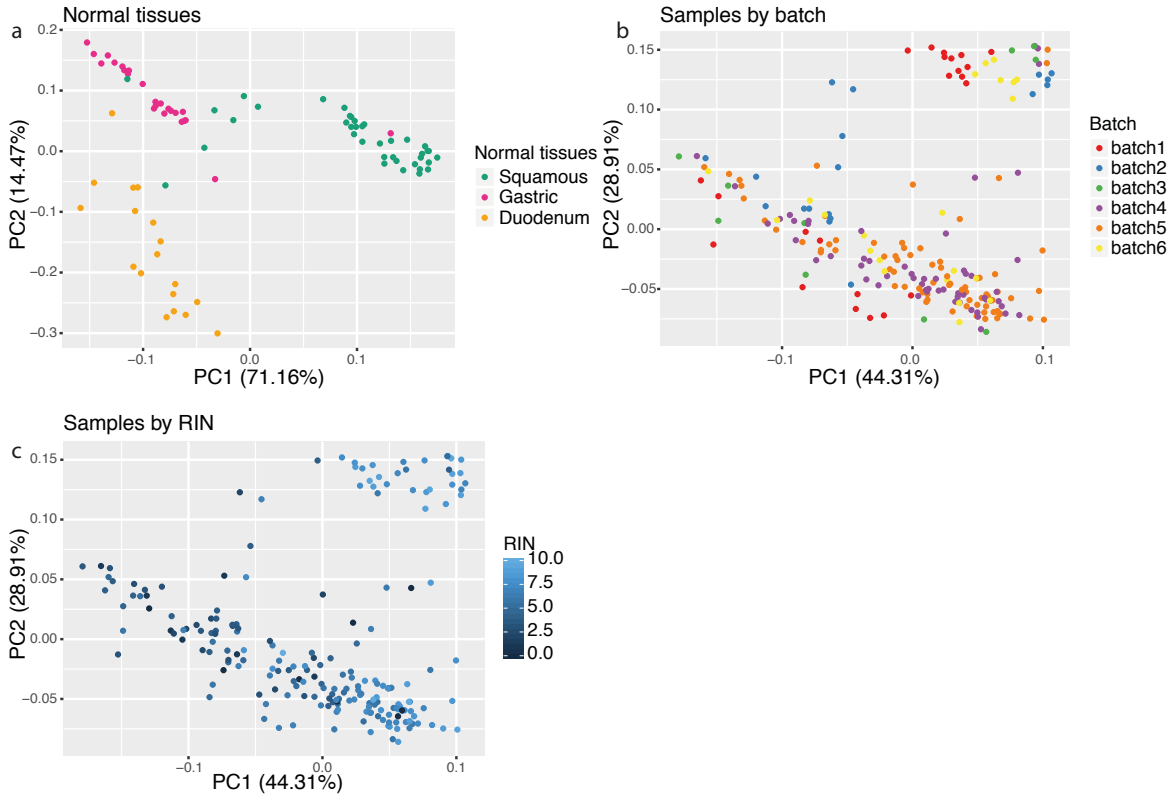


Figure 34 Principal component analysis of all normal tissue and all samples by batch and RIN

a. All normal tissues. Percentage of variability explained by each principal component (PC) displayed in parentheses. **b.** All samples coloured by batch. **c.** All samples coloured by RNA Integrity Number (RIN).

On considering all the samples, the squamous phenotype contributed predominantly to principal component (PC) 2 (Figure 35a). This was not surprising as squamous epithelium has a very different expression profile to the glandular epithelium of BE, gastric and duodenum. In order to better observe the separation of the other groups, we repeated the PCA excluding the squamous. BE samples, regardless of dysplasia status, formed a distinctive group which, surprisingly, was distinct from duodenum with PC2. There was some overlap of BE with gastric samples but it was mainly the Trio BE rather than the pre-cancer BE (Figure 35b,c). This could be because they contained gastric metaplasia (GM), whereas we were careful to exclude samples with GM when creating the pre-cancer BE cohort.

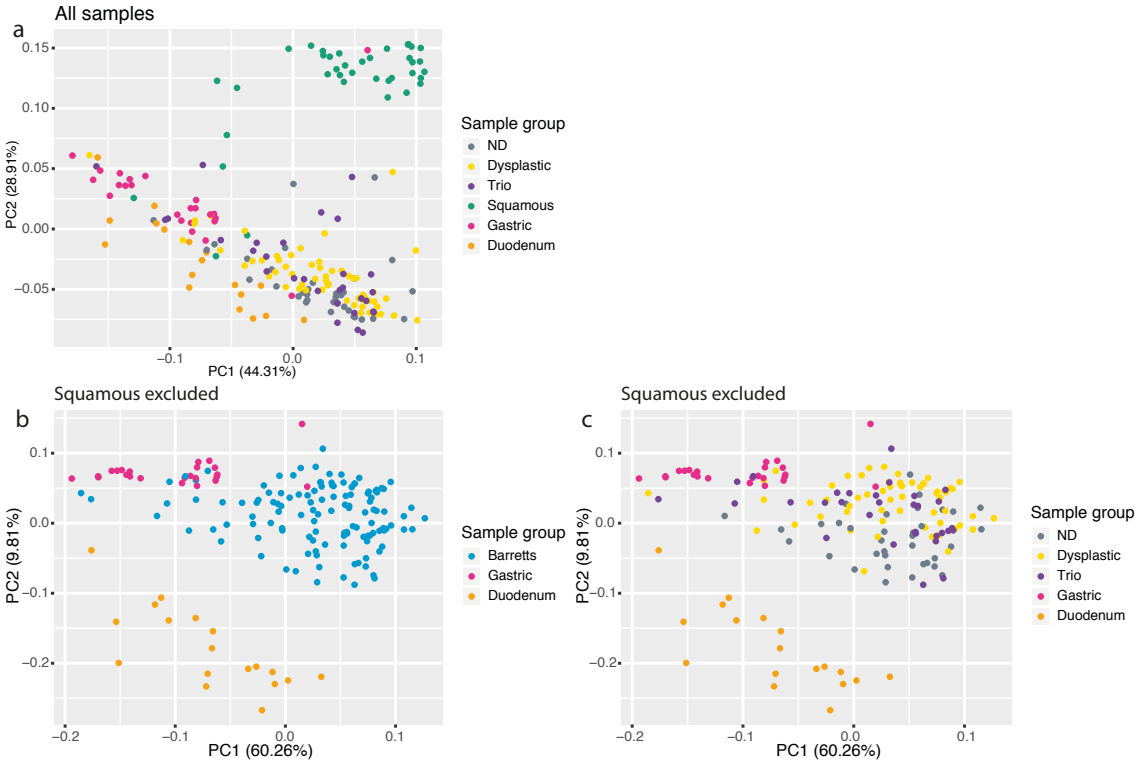


Figure 35 Principal component analysis of all samples

a. All samples coloured by group. **b.** Squamous excluded from the analysis, coloured by overall group. **c** Same plot as in **b.** with BE samples further coded by grade. ND = non-dysplastic. PC = principal component.

Next we focussed in on how well the pre-cancer ND and dysplastic BE would cluster independently. PCA of the 1000 most variable genes between ND and dysplastic BE separated the pre-cancer BE cohort well on PC2 but not so much so on PC1 (Figure 36a). This was surprising however may possibly be explained by the overall composition of frozen biopsies and the presence of inter-mixed ND epithelium within dysplastic biopsies. Whilst the highest cellularity biopsies possible were used, this would be unavoidable without microdissection.

We used the same 1000 genes to compare the Trio BE to the pre-cancer cohort. The Trio BE scattered throughout the pre-cancer cohort, rather than clustering independently (Figure 36b). When coloured for their grade many of the Trio BE that had been graded pathologically as ND clustered with the pre-cancer dysplastic samples (Figure 36c). This finding was consistent with the genomic patterns seen in Results 1. It seems that although pathologically, BE adjacent to cancer looks like ND non-progressing BE, neither the genomic nor expression profiles convey this.

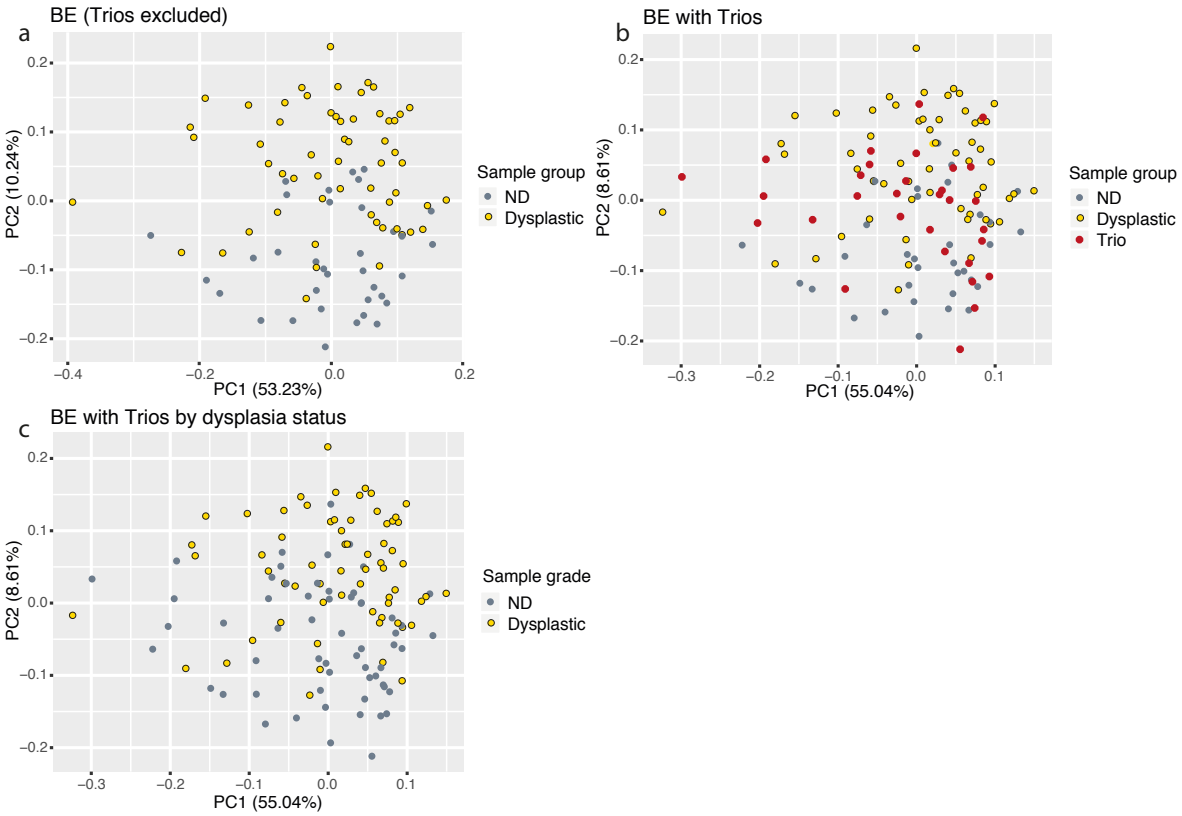


Figure 36 Principal component analysis of Barrett's oesophagus samples

a. Trio BE samples excluded. Colour coded by overall dysplasia status. **b.** Trio BE included. **c.** Trio BE included but coded by their dysplasia status. ND = non-dysplastic.

4.2 Differential gene expression analysis

4.2.1 Expression in dysplastic versus non-dysplastic

In order to understand the changes in gene expression with progression we compared the ND and dysplastic samples of the pre-cancer BE cohort by performing a differential analysis using DESeq2 (Love et al., 2014b). When considering a differential expression of greater than 3-fold (log2 fold change (FC) > 1.58) there were 352 significantly upregulated genes and 123 significantly downregulated (q value < 0.05) in dysplastic versus ND samples (Supplementary table 1) (Figure 37). Given the heterogeneity of BE, a fold change of 3 was chosen to avoid small magnitude changes that would unlikely be biologically significant.

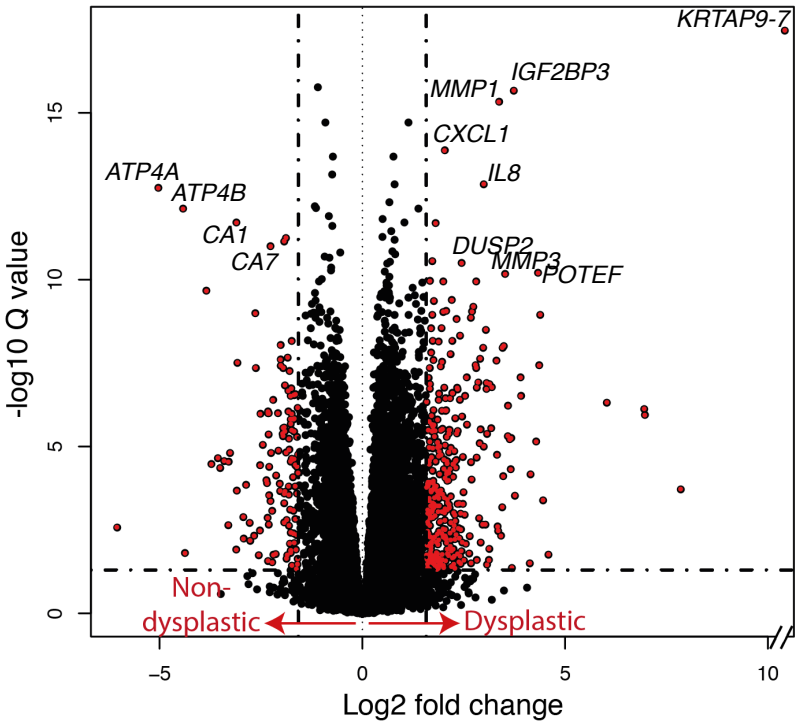


Figure 37 Volcano plot of differentially expressed genes between dysplastic and non-dysplastic

Genes with expression greater than 3-fold change difference (log2FC > 1.58 or <-1.58) and q value < 0.05 coloured in red. Vertical dashed lines represent the FC cut offs. Horizontal dashed line marks the q-value threshold for significance. Outlier genes individually labelled.

The extreme log2FC of 30 in *KRTAP9-7* was driven by a very high expression in one sample. (530 vs. a median of 0 for other samples). It was expressed in one other dysplastic case at a low level but not in any other dysplastic or non-dysplastic cases. This one high count was enough to contribute to the significant fold change which DESeq2 computes. *KRTAP9-7* is a keratin-associated protein only expressed in skin, (http://www.uniprot.org/uniprot/A8MTY7#function). It was not expressed in the squamous samples in our cohort and it is not expressed in other cancers (http://www.proteinatlas.org/ENSG00000212659-KRTAP9-6/pathology). Given that it was only expressed in one sample it was not considered to be relevant and likely from contamination.

In order to understand more about the biological processes in progression, rather than focussing on specific outliers, we looked at the gene deregulation as a whole and the pathways involved.

4.2.2 Significantly deregulated genes in dysplasia

Using the significantly up- and downregulated genes between ND and dysplastic, from the DESeq2 output, we compared the expression across the samples. Firstly, we considered only the pre-cancer cohort and found the ND and dysplastic cases to cluster distinctly with hierarchical clustering (Figure 38). However, there was no distinct clustering of the grades (non-progressors (NP), pre-progressors (PP), LGD, HGD, IMC) either by taking the patient grade or the dominant grade in the biopsy. Although HGD (orange) and IMC (red) did dominate the right of the hierarchical clustering dendrogram. The NP (blue) and PP (purple) samples were indistinguishable, although this is not necessarily surprising as one would not expect expression changes to occur so far in advance of phenotypic change and they were genomically similar in Results 1.

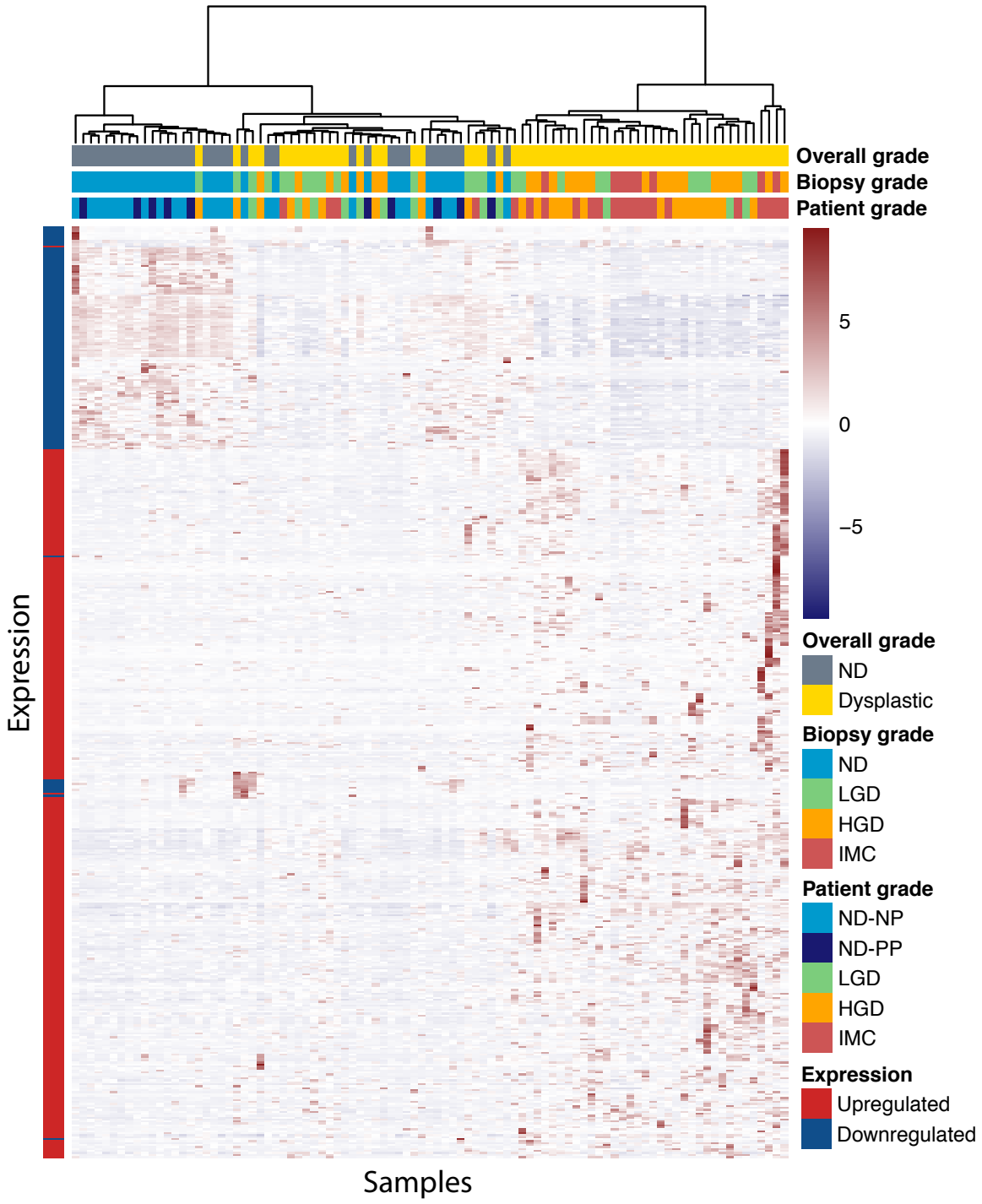


Figure 38 Clustered heatmap of expression of up or downregulated genes in the pre-cancer Barrett's oesophagus cohort

Each row represents a significantly up or downregulated gene (in dysplasia compared to ND) from the DESeq2 output. The columns are samples. Expression is scaled for each gene using the Z-score (subtracting the mean and dividing by the standard deviation). High expression is red, low expression is blue. The dendrogram at the top shows the hierarchical clustering (using the ward D method) of the samples. The horizontal annotation bars show the overall grade of the patient, the biopsy subgrade and the patient subgrade. ND = non-dysplastic, ND-NP = non-dysplastic non-progressor, ND-PP = non-dysplastic pre-progressor, LGD = low grade dysplasia, HGD = high-grade dysplasia, IMC = intramucosal carcinoma.

In Results 1, there was no significant difference in the median numbers of mutations, copy number aberrations or structural variants between the Trio BE and the pre-cancer BE. In order to see if there was any difference at the expression level, we added the expression of the above genes in the Trio BE. We found that the Trio BE samples were a mix of the patterns seen in pre-cancer ND and dysplastic samples, irrespective of the grade of the Trio BE (Figure 39). ND Trio BE both genomically and transcriptomically behaved more like dysplasia despite its indolent phenotypic appearance. Some Trio BE cases appeared to have upregulation of dysplasia genes, whilst still expressing the ND genes. This strengthened the finding we had seen in the earlier PCA and also in the genomic analysis in Results 1.

The Trio BE were labelled as dysplastic if they pathologically had any dysplastic cells at all in the frozen specimen. Thus, the percentage of dysplasia was generally lower than in the pre-cancer cohort. This could explain the expression of ND genes in the dysplastic Trio BE, but not the opposite: seeing the expression of the dysplasia genes in the ND samples. So, the ND Trio BE appears to have a different expression profile which is in a hybrid state between ND and dysplastic cases.

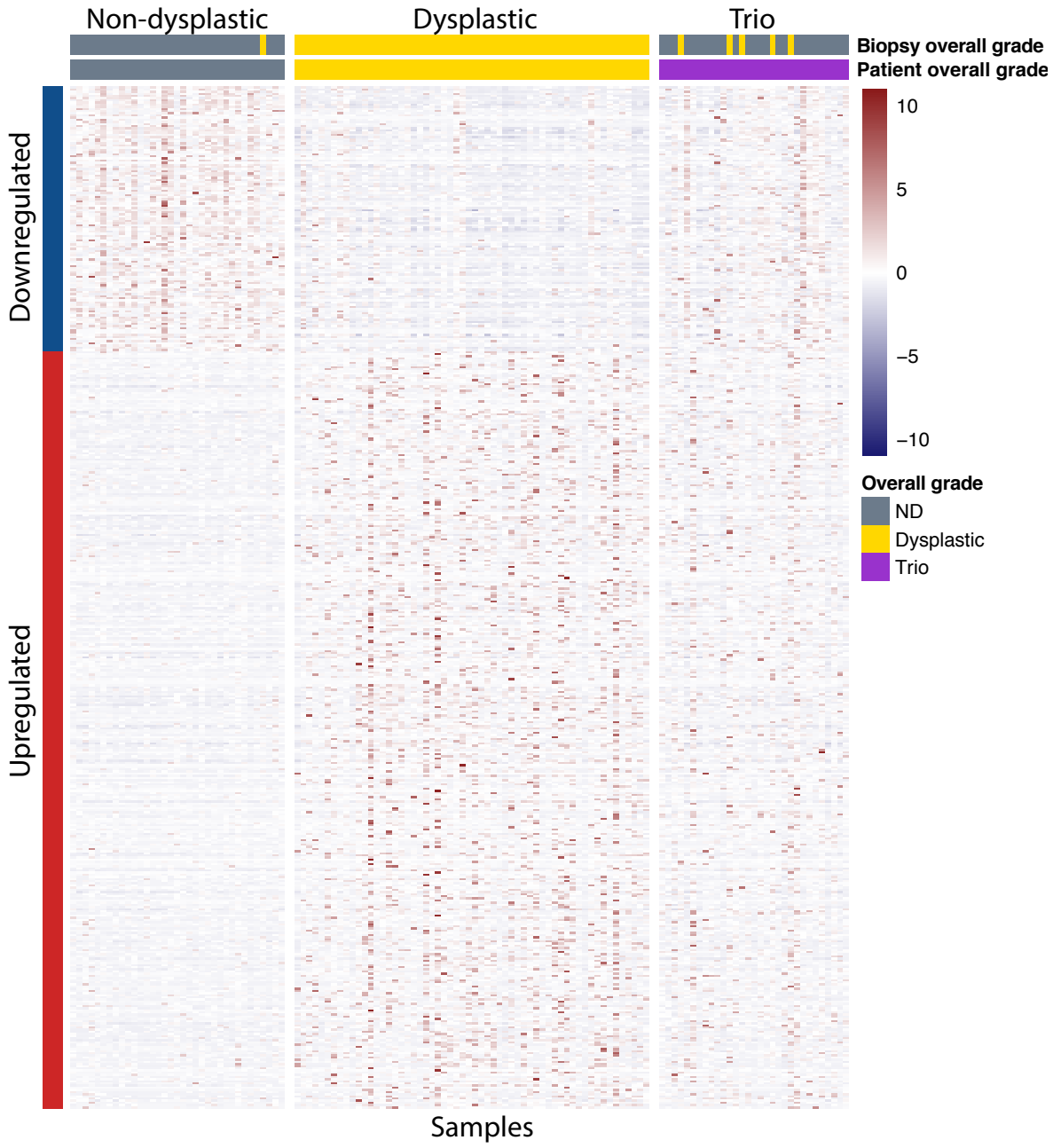


Figure 39 Unclustered heatmap of expression of up or downregulated genes in the pre-cancer Barrett's oesophagus cohort compared to the Trio BE.

Each row represents a significantly up or downregulated gene (in dysplasia compared to ND) from the DESeq2 output. The columns are samples ordered by grade. Expression is scaled for each gene using the Z-score. High expression is red, low expression is blue. The horizontal annotation bars show the overall grade of the patient or the biopsy grade. Grade of the Trio BE samples is coded in the second horizontal annotation bar. ND = non-dysplastic.

Lastly, we took the significantly up and downregulated genes and compared their expression in the pre-cancer BE samples to the normal squamous oesophagus, gastric cardia and duodenum in order to ascertain whether they were expressed in normal tissue or specific to BE (Figure 40). This differed to the prior comparison in Figure 35 of the most variably expressed genes across all tissue types. Some of the differentially expressed genes were strongly expressed in squamous tissue. These genes were expressed in a small number of BE samples and this was likely due to squamous contamination. However, this was enough to drive the significant expression differences seen in the analysis.

Excluding the squamous genes, the genes expressed in the ND samples (downregulated in dysplasia) were also highly expressed in duodenum (Figure 40). The intestinal metaplasia of Barrett's oesophagus has a phenotype similar to duodenal tissue, with pathognomonic goblet cells, and so this would not be unexpected. However, these genes were not expressed in the dysplastic samples, suggesting some loss of this intestinal phenotype with progression. In contrast, the genes upregulated in dysplasia were not strongly expressed in any of the normal control tissues.

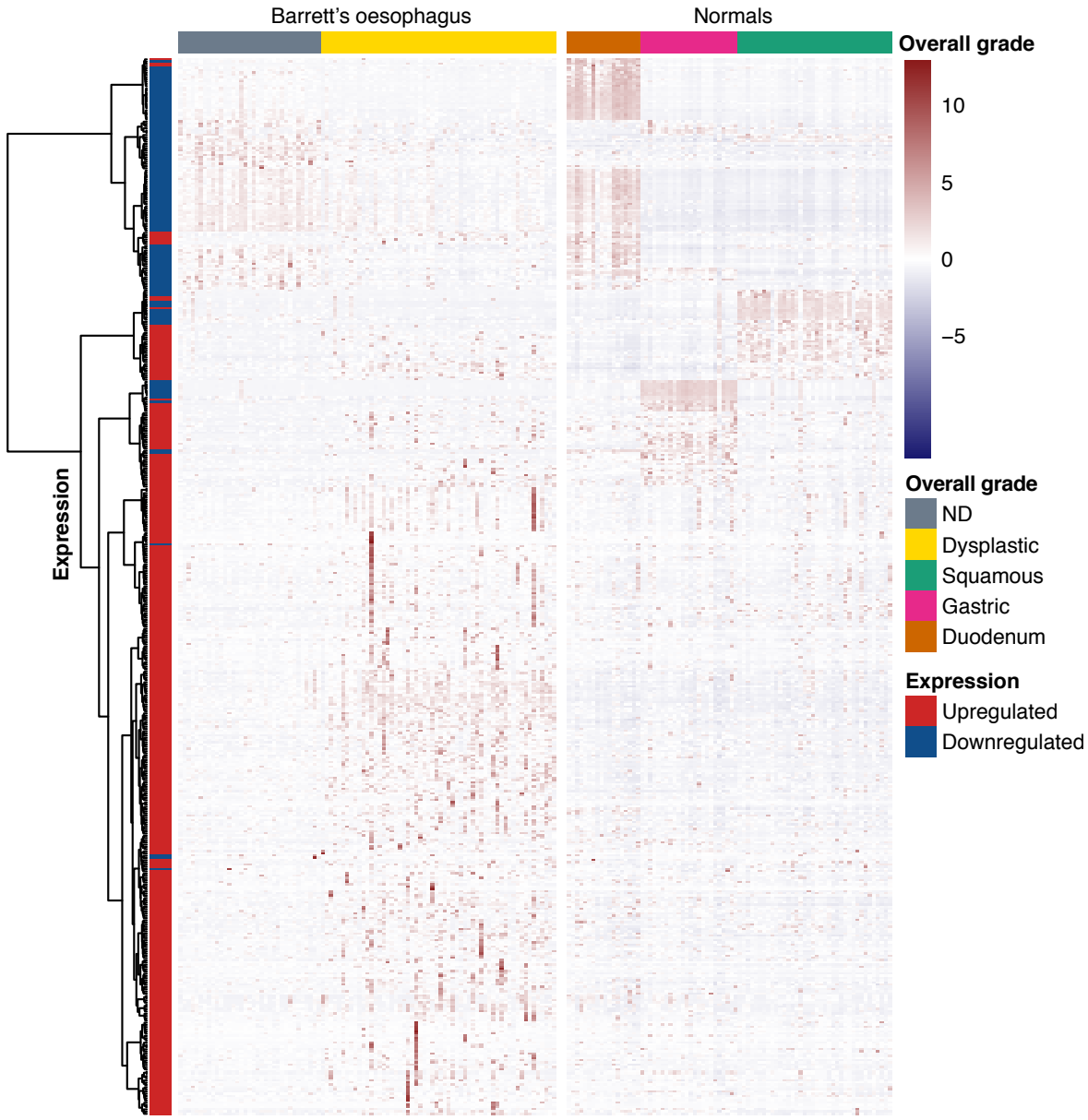


Figure 40 Heatmap of pre-cancer Barrett's oesophagus samples compared to the normal tissue

Each row represents a significantly up or downregulated gene (in dysplasia compared to ND) from the DESeq2 output. The rows are clustered hierarchically using the ward D method. The columns are samples ordered by grade and normal tissue type: duodenum, gastric, squamous. Expression is scaled for each gene using the Z-score. High expression is red, low expression is blue.

To further evaluate this, we took all the genes with upregulation of expression in both NDBE and duodenum from the analysis (82 genes) and compared the overall expression of these genes between the grades and all tissue types.

The mean Z-score for these genes was calculated for each patient and these means compared for each group. Figure 41 shows that the genes expressed in duodenum are also expressed in NDBE, but that this expression is significantly lower in dysplastic tissue ($p=5.1x10^{-13}$, Wilcoxon Rank Sum). As expected, these genes are hardly expressed at all in gastric and squamous samples. In the earlier PCA, in Figure 35, the most variable genes between all samples were considered and, overall, the BE was more similar to gastric. Instead, this differential analysis was considering only the differential genes between dysplastic and ND, rather than tissue similarity. So, the genes which are downregulated in the transition from ND to dysplastic are predominantly also expressed in duodenum.

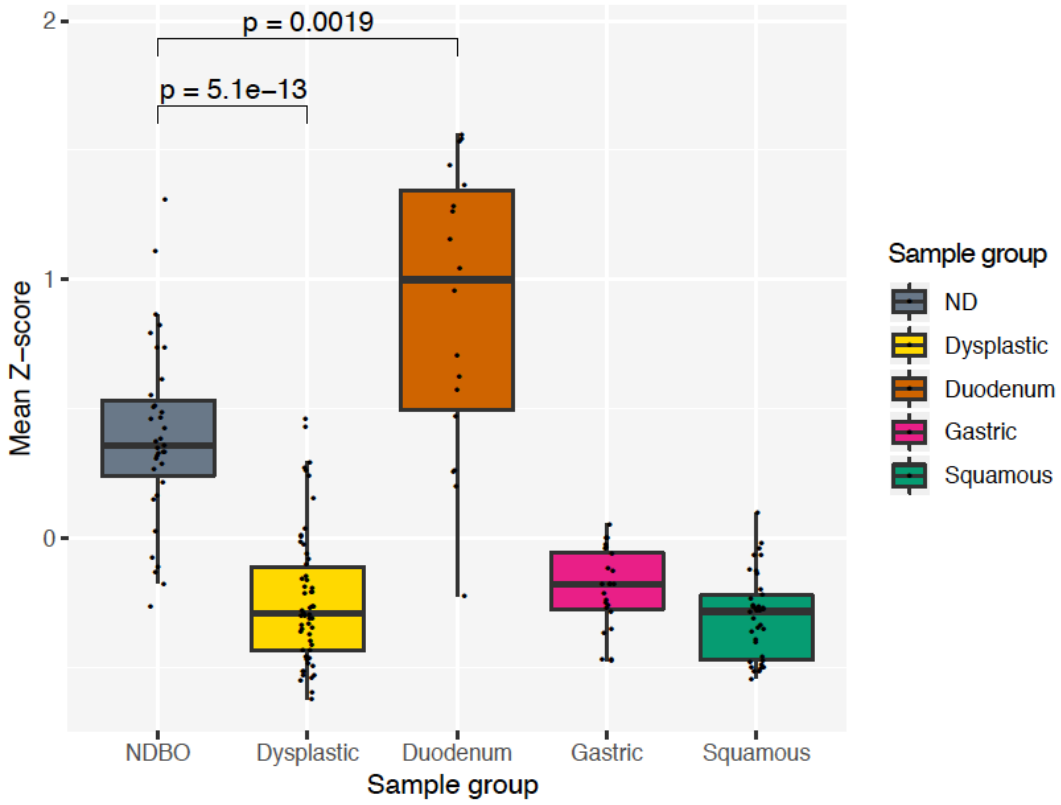


Figure 41 Downregulation of intestinal phenotype with progression

Boxplot of mean expression of genes upregulated in duodenum per sample. All genes were taken from the differential expression analysis that were expressed highly in duodenum. Z-score calculated for every sample for each gene and a mean was taken of the Z-scores of all genes per sample. This gave a proxy value for the overall expression of duodenal genes for each sample. Mean values plotted as a boxplot for each group. P-values calculated using Wilcoxon Rank Sum test.

In order to understand more about the functions of the downregulated genes, we ran Ingenuity Pathway Analysis (IPA; Qiagen, Germany) and stringDB (https://string-db.org). The most significant Gene Ontology terms for molecular and cellular functions were lipid metabolism and molecular transport, digestion and intestinal absorption (e.g. *CDX1*). *APOBEC1*, involved in the editing of cytosine to uracil and usually only expressed in small intestine, was in the cluster. The top predicted regulator effect network included *HNF4A* and *HNF1A*. These are transcription factors which are thought to be important in the development of the intestines, liver and kidney (https://www.ncbi.nlm.nih.gov/gene?cmd=Retrieve&dopt =full_report&list_uids=3172). IPA also showed them to interact with three other upstream regulators, resulting in a complex network of gene regulation involved with glycolysis and transport of molecules (Figure 42). *PPARG* encodes a nuclear receptor with roles including the controlling the peroxisomal beta oxidation pathway of fatty acids and gut homeostasis via suppressing NF-kappa-B-mediated proinflammatory responses (http://www.uniprot.org /uniprot/P37231#function). A number of the genes e.g. *APOB, MTTP, FABP1, FABP2* are involved specifically in lipid transport, a main function of the small intestine. The loss of genes involved in lipid transport is consistent with the loss of phenotype seen with progression.

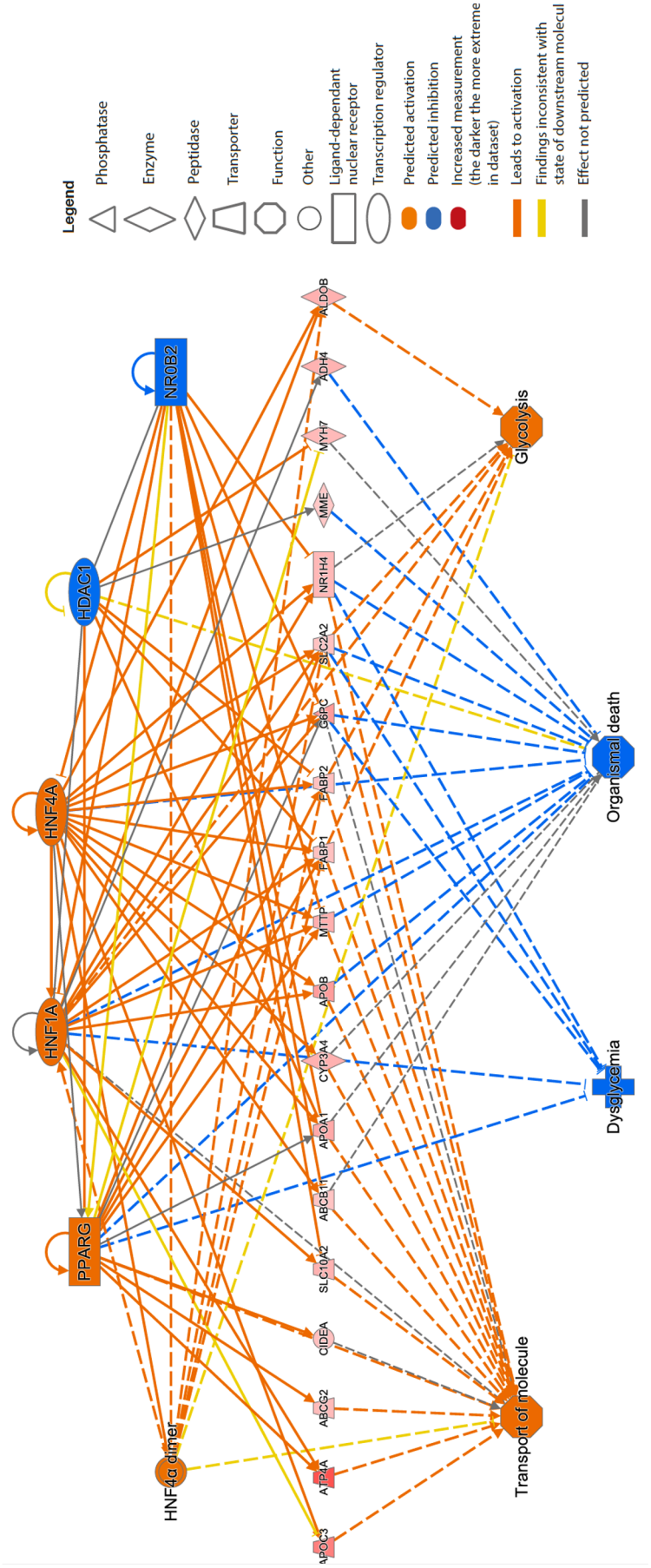


Figure 42 Ingenuity Pathway analysis of upstream regulation of genes upregulated in non-dysplastic Barrett's oesophagus

IPA identified the interaction of 6 upstream regulators in the genes controlling 19 genes which were significantly expressed in non-dysplastic Barrett's oesophagus compared to dysplastic.

MUC2, encoding the secretory protein mucin-2, was significantly down-regulated from ND to dysplasia. It is a goblet cell marker specific to duodenum, small intestine, colon and rectum (https://www.proteinatlas.org/ENSG00000198788-MUC2/tissue). It is known to be a marker for intestinal metaplasia in BE (Lavery et al., 2014; Reis et al., 1999; Zhou et al., 2019b). We separately considered the expression of 2 other genes known to be goblet cell markers: *TFF3* and *REG4*. They were not found to be differentially expressed in the DESeq2 analysis because there was a <3-fold difference between the dysplastic and ND. However, all three showed a significant reduction in expression with progression, with very low expression in gastric and squamous. These findings are also in keeping with the concept of a loss of intestinal phenotype with progression.

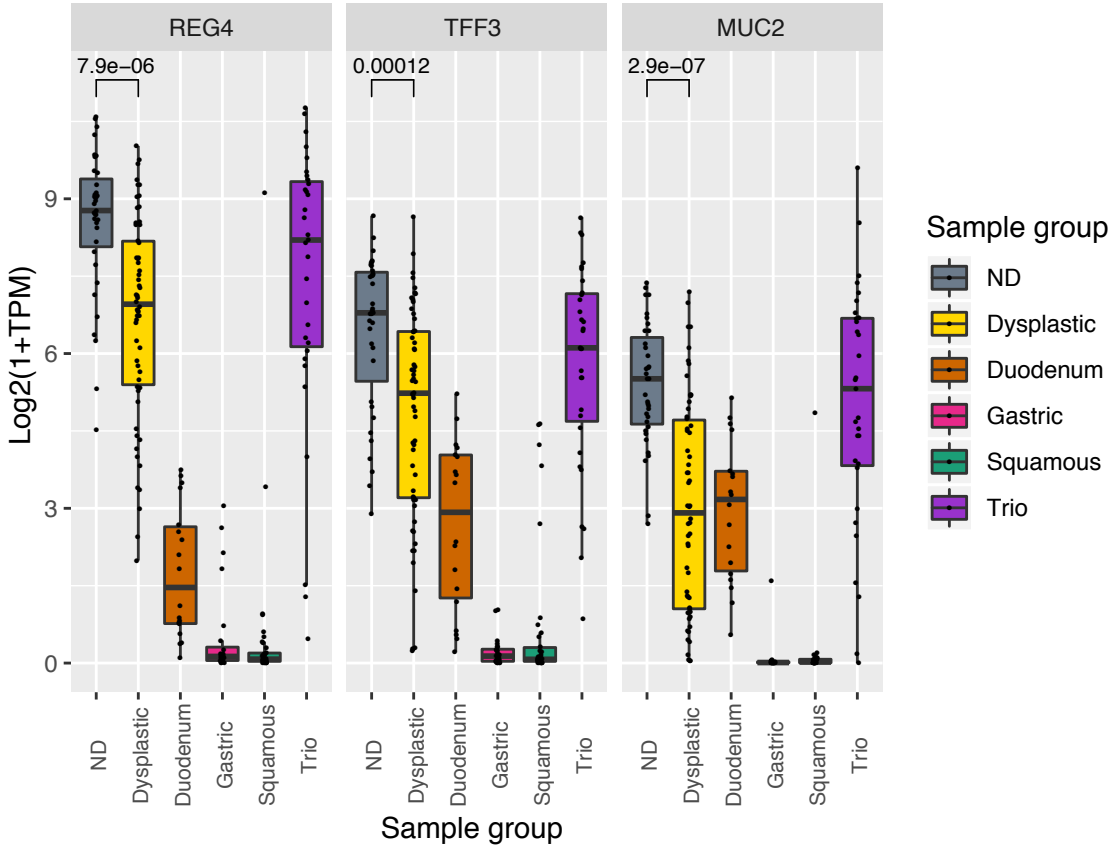


Figure 43 Expression of goblet cell markers in different tissue types

Boxplots showing expression (log2(1+TPM)) of goblet cell markers in non-dysplastic (ND) versus dysplastic Barrett's oesophagus compared to normal tissues. Trio = BE adjacent to cancer. P-values given between ND and dysplastic (Wilcoxon Rank Sum test). TPM = transcripts per kilobase million.

Overall, the genes which are downregulated in the progression from ND to dysplastic, were found to be normally expressed in duodenum but not gastric or squamous. They are predominantly involved in normal intestinal functions including absorption, secretion and lipid metabolism. On histology, we see a loss of intestinal metaplasia with progression (Naini et al., 2016; Odze, 2006) and the expression data supports this.

4.3 Pathway analysis of differentially-expressed genes in pre-cancer dysplasia

Next we focussed on the genes upregulated in the dysplastic samples. In order to see if these genes grouped into pathways, we again used Ingenuity Pathway Analysis (IPA). All the significantly up and downregulated genes in dysplasia, with a fold change >3 were inputted. An upstream analysis identified 126 of the genes as being downstream of the mitogen-activated protein kinase (MAPK/ERK) pathway. The MAPK pathways have long been known to have roles in the regulation of cell proliferation, apoptosis and differentiation (reviewed in (Burotto et al., 2014; ZHANG and LIU, 2002)). These genes and their interactions are shown in Figure 44. Some of the downstream genes in the pathway, e.g. *CXCL1*, *CXCL2* and *IL8* encode chemokines or other genes important in the immune response (highlighted in pink in Figure 44). CXCL1 is a chemoattractant of neutrophils in inflammation and CXCL2 is produced by activated monocytes and neutrophils. *IL8* (also called *CXCL8*) encodes a chemokine secreted by macrophages. It was not possible to say if there was upregulation of these genes because of the MAPK pathway or because there was increased inflammation. However, the pathology reports of the biopsies did not indicate a difference in inflammation between the dysplastic and the ND samples. Alternatively, the immune environment may independently facilitate progression, which is further explored later in this chapter.

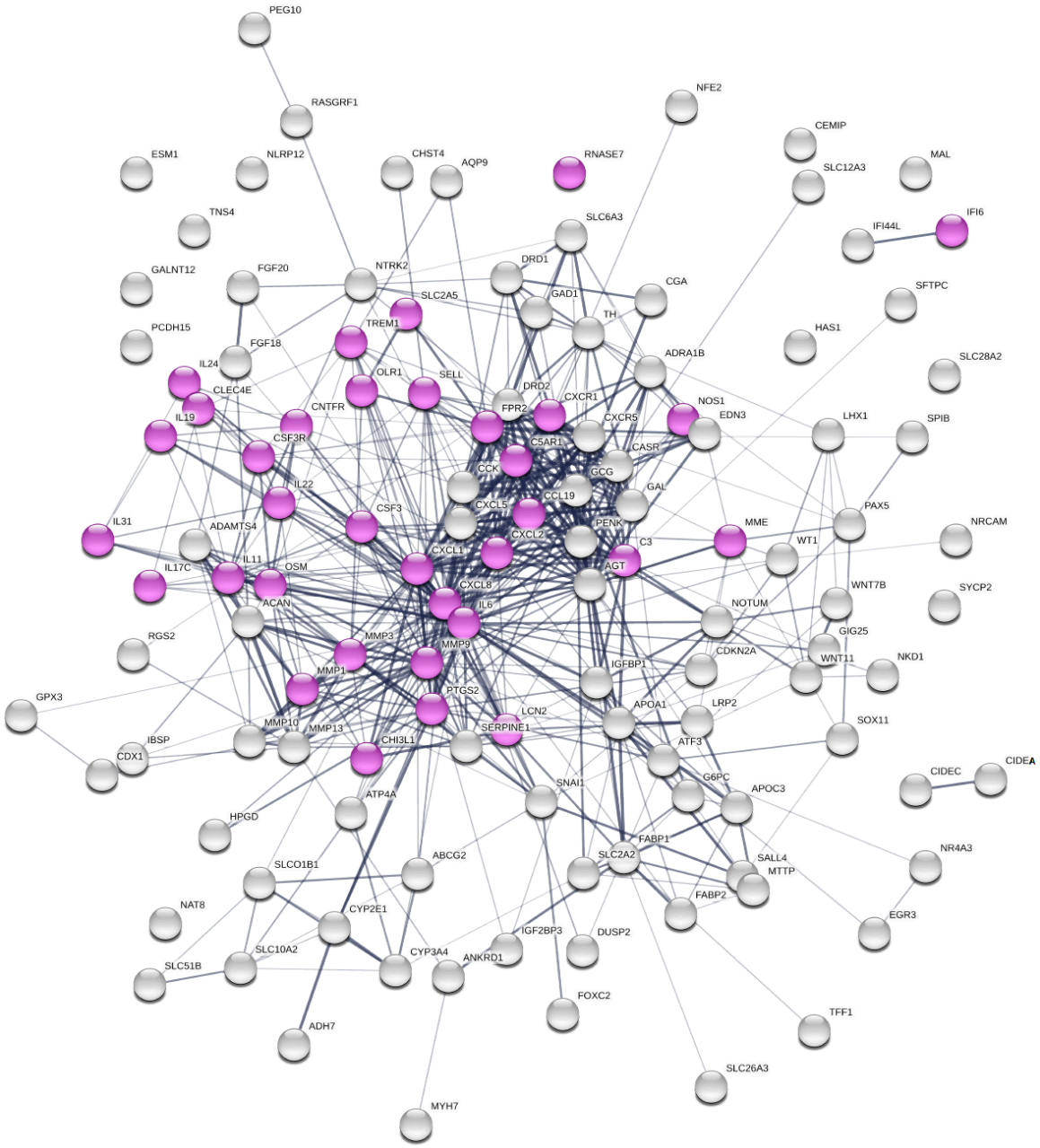


Figure 44 Interactions of all genes downstream of the ERK network

Visual representation of the interactions between the 126 genes altered in my cohort which are downstream of regulators in the ERK network. Genes highlighted in pink are also in the 'Immune System' Reactome Pathway. Intensity of grey line represents strength of data support. Generated using the STRING database (https://string-db.org).

Fifteen of these genes are directly regulated by ERK1/2 (Figure 45a). The mechanistic network function in IPA predicted the activation or inhibition of 20 upstream regulators of these genes (Figure 45b).

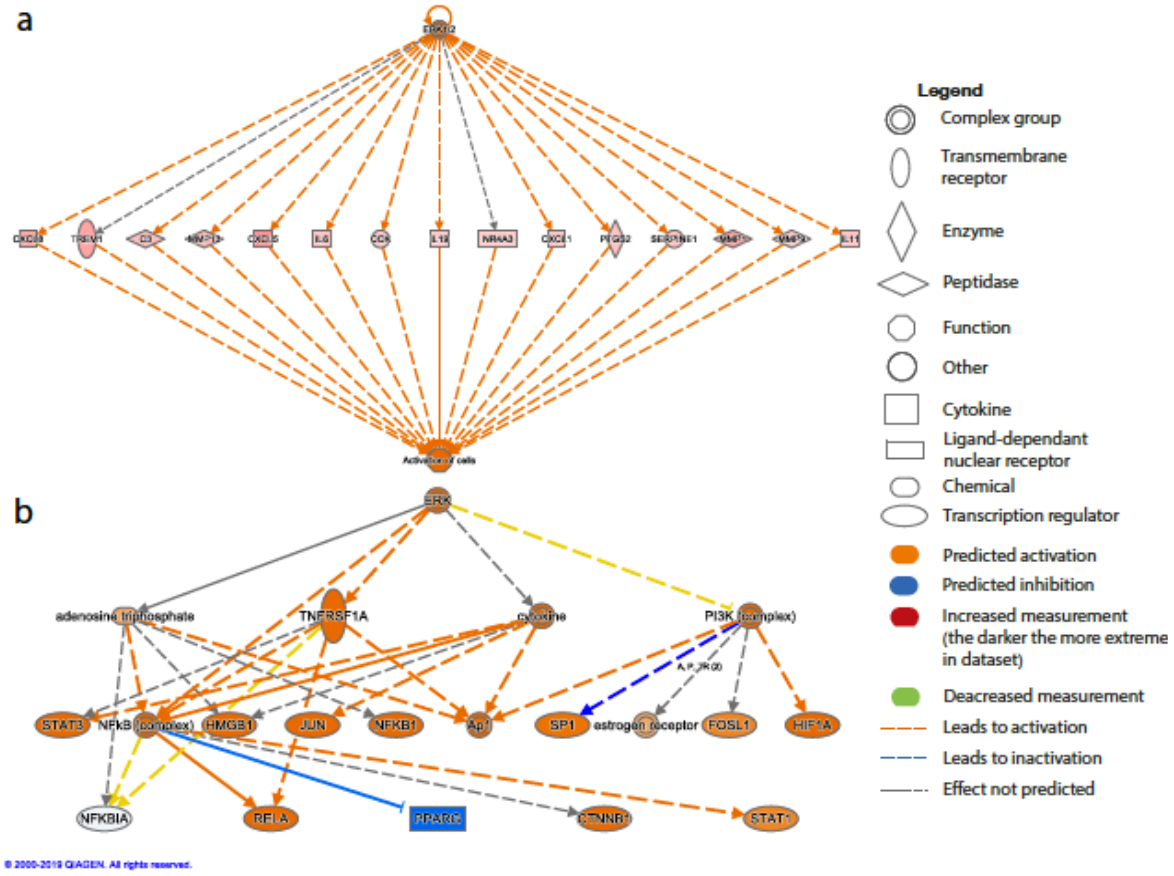


Figure 45 The MAPK pathway

a. Fifteen genes directly regulated by ERK. Genes shaded in red are upregulated in our analysis. The shape of the icon represents protein function. **b.** Ingenuity Pathway Analysis prediction of all upstream regulators in the MAPK pathway. Upstream of the 126 genes in Figure 44. Genes predicted to be activated or inhibited based on the expression of the downstream genes in the samples in the Barrett's oesophagus cohort. Predicted activation = orange, predicted inhibition = blue.

This analysis indicates upregulation of the MAPK pathway in dysplasia. However, this effect could have been driven by high expression in a small number of samples. To consider this, a heatmap was plotted of the expression of just these 126 genes. In Figure 46a the samples are sorted by patient grade but the genes in the rows are hierarchically clustered. 89 of the 126 genes were upregulated in the differential analysis, indicated in red in the first vertical annotation bar, and the rest were downregulated. The vertical annotation bar gives the IPA prediction as to whether the observed expression would be due to activation or inhibition of the upstream pathway. The majority of the labels are green in the upregulated genes, indicating overall pathway activation. The predictions are given by IPA, using inbuilt literature review information, and should only be used as a guide. They are also simplified as only the dominant prediction was used across multiple regulators. The heatmap shows that the upregulation does occur across all the dysplastic samples and is not seen in the ND. In contrast, there is upregulation of a small number of genes in the ND which inhibit the pathway and have loss of expression with progression.

To strengthen these findings, we looked at the overall upregulation of genes in the pathway per sample and compared each grade. To do this we took only the upregulated genes and calculated the mean Z-score for each sample as a proxy for the overall activation. Figure 46b shows a significant overall upregulation of the pathway in the dysplastic cases ($p = 4.5x10^{-12}$, Wilcoxon Rank Sum test). There was a significant increase from ND-PP to LGD, but no significant difference between the dysplastic grades (Figure 46c).

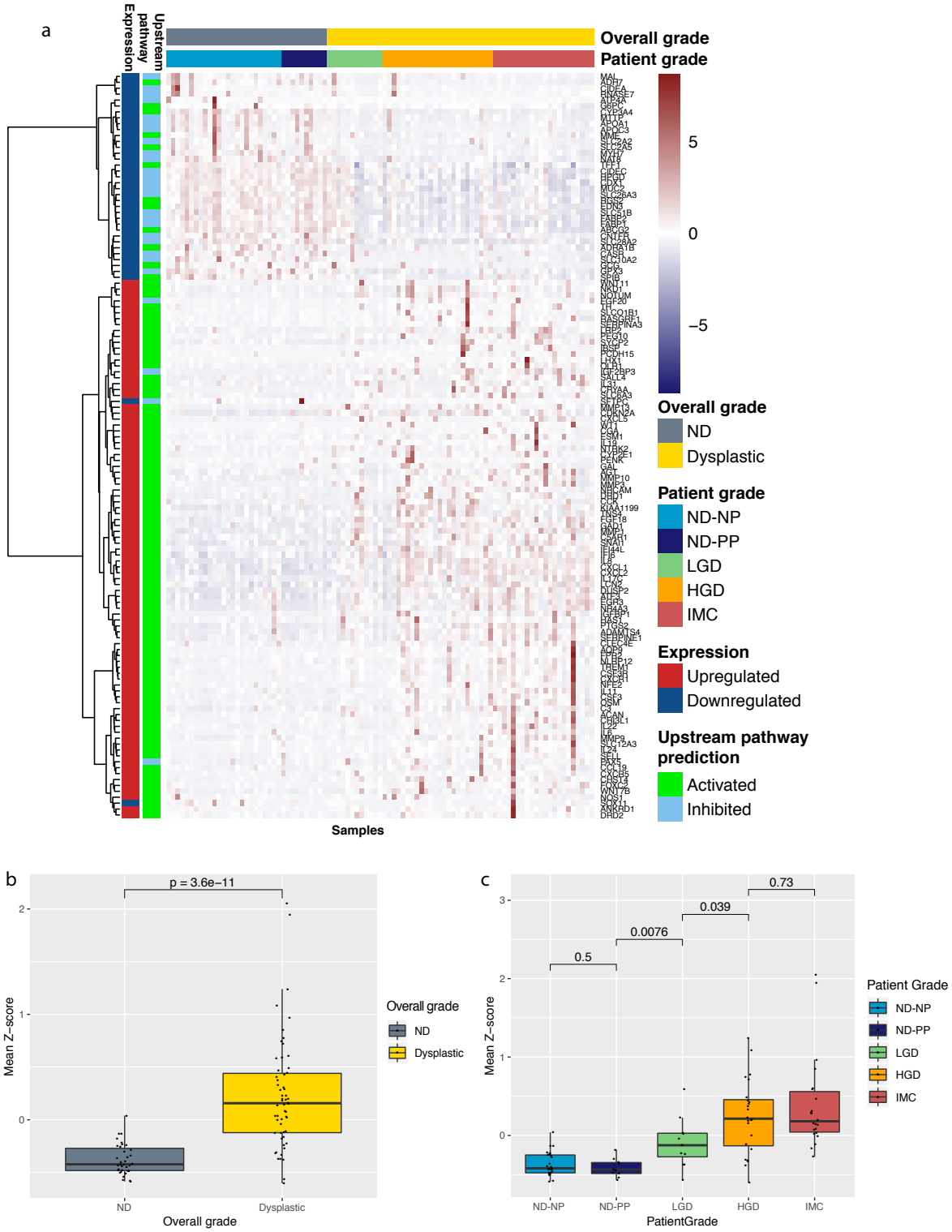


Figure 46 Expression of genes downstream in the ERK/MAPK pathway

a. Heatmap of the scaled expression of the 126 genes downstream of ERK in the non-dysplastic (ND) and dysplastic Barrett's oesophagus. Clustering of genes in the rows (ward D). Rows scaled using Z-score. Samples in the columns ordered by grade. First vertical annotation bar indicates if the gene was up (red) or down (blue) regulated in dysplasia in the differential analysis. The second bar is the Ingenuity Pathway Analysis prediction of whether the expression seen would lead to activation (green) or inhibition (blue) of the MAPK pathway. **b. c**. Boxplot representation of the overall MAPK pathway upregulation. Z-scores calculated for each gene across the samples. A mean Z-score was then calculated for each sample as a proxy for the overall representation of the pathway in the sample. Mean scores plotted per group. P values calculated between groups (Wilcoxon Rank Sum test). ND-NP = non-dysplastic non-progressor, ND-PP = non-dysplastic pre-progressor, LGD = low grade dysplasia, HGD = high-grade dysplasia, IMC = intramucosal carcinoma.

Gene set enrichment analysis corroborated the above findings, with enrichment of genes in the AP1 (FOS/JUN) pathway (Schaefer et al., 2009) that acts immediately downstream of ERK1/2 (Figure 47).

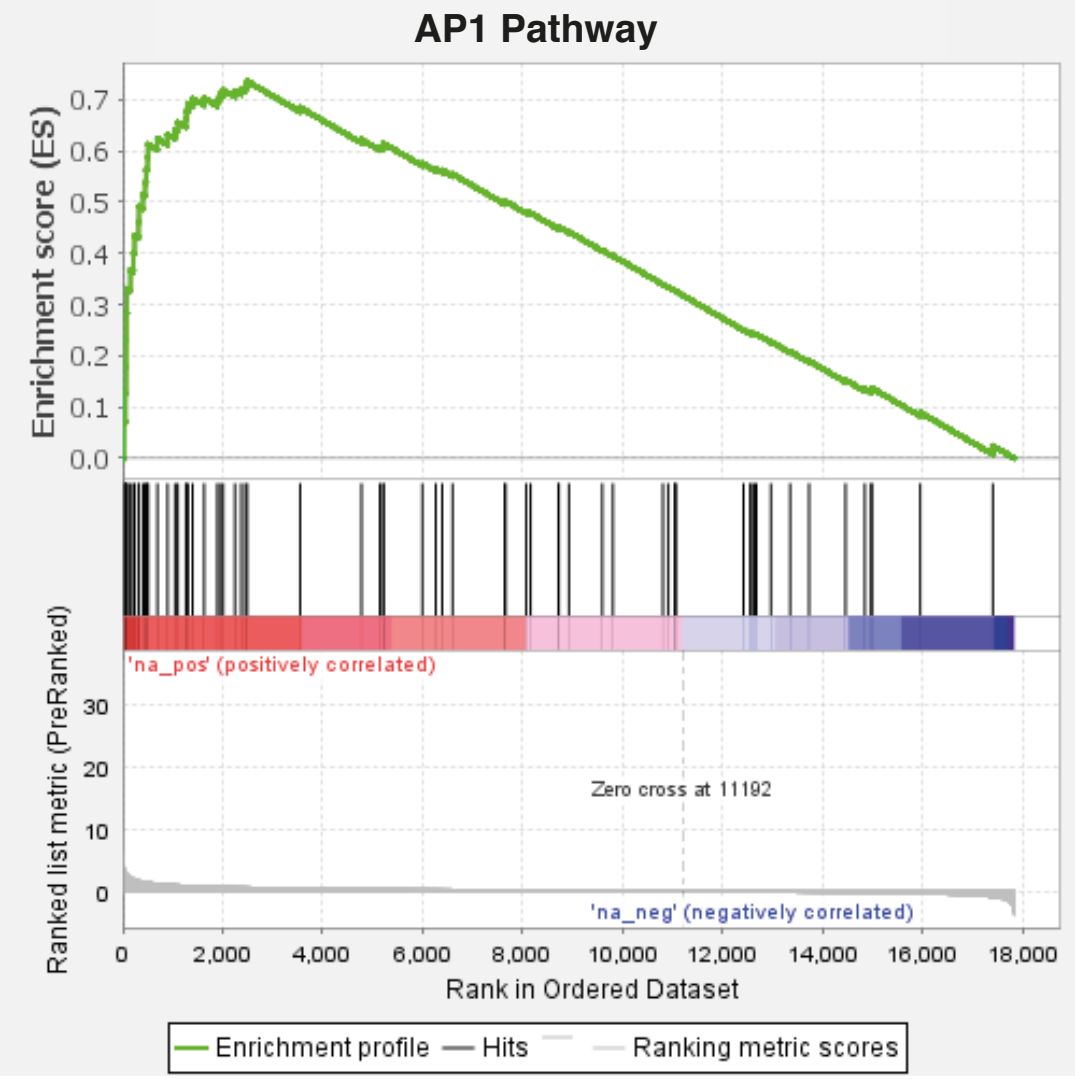


Figure 47 Gene set enrichment analysis of the AP1 pathway

AP1 (FOS/JUN dimer) is activated by ERK in the MAPK pathway. All genes from the differential analysis used with their calculated fold changes. List of genes curated by GSEA based on the literature (Schaefer et al., 2009).

This analysis only considered genes which were found to be significantly deregulated in dysplasia vs. ND in the DeSeq2 analysis. It is important to consider any genes in the pathway which were not deregulated. Furthermore, many of the genes identified as being downstream in the network are also downstream in other pathways. We used an independent published, curated list (Ünal et al., 2017) of direct targets of ERK and considered the expression of these 246 genes in the pre-cancer cohort using hierarchical clustering Figure 48.

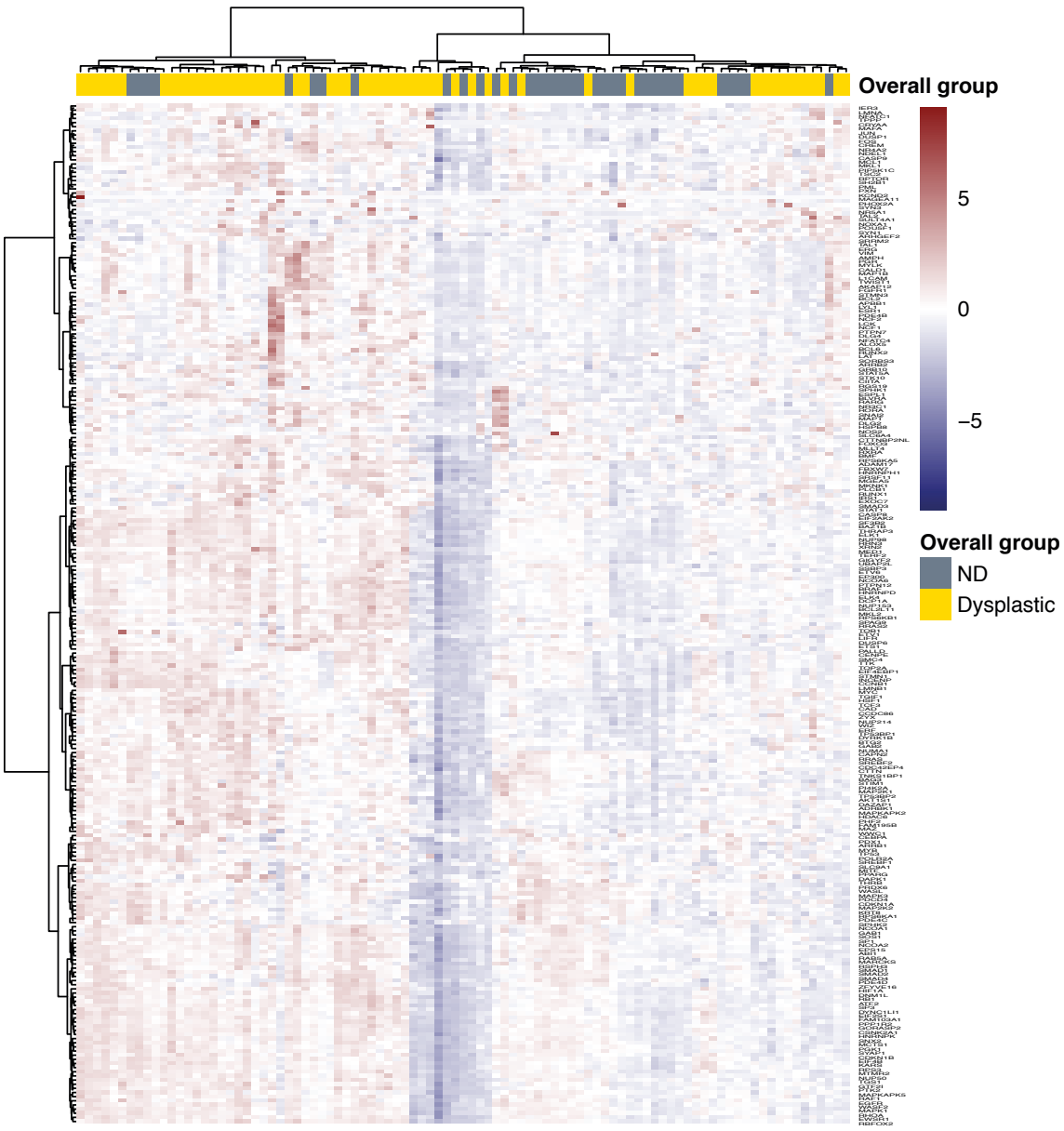


Figure 48 Expression of direct downstream targets of ERK

Heatmap of gene expression, in the pre-cancer cohort, of a published, curated list of 248 direct downstream targets of ERK. Hierarchical clustering of non-dysplastic (ND) and dysplastic Barrett's oesophagus in columns, and genes in the rows (ward D). Rows scaled using Z-score.

The heatmap shows increased expression of the genes in the main cluster of dysplastic samples, on the left of the heatmap. The differential expression is less clear as with the earlier gene set from IPA, but that is expected given that the previous gene set was sampled from genes that had been confirmed to be differentially expressed. Overall this analysis supports the above findings, that there is overall upregulation of the ERK/MAPK pathway in dysplasia.

4.4 Immune infiltration in progression

4.4.1 Introduction

With the fast expansion of large data sets of bulk RNA sequencing, computational methods have been developed to analyse the complete immune infiltrate. Gene Set Variation Analysis (GSVA) is a tool which assigns immune enrichment scores to samples based on the expression of different immune markers (Hänzelmann et al., 2013). This allows the deconvolution of the expression data to predict the proportions of each immune cell type within the sample. We were interested to look at the composition of the immune microenvironment as it is becoming increasingly clear in cancers that the tumour microenvironment (TME) can impair anti-tumour immunity by polarising the immune composition towards less cytotoxic subsets e.g. T regulatory cells (Lin et al., 2016). T regulatory cells have a suppressive role in normal physiology by regulating the activation and expansion of T and B cells. In the tumour setting this can result in reduced anti-tumour immunity due to inhibiting cytotoxic T cells and the secretion of immunosuppressive cytokines. Increased cytotoxic T cell infiltrations are associated with an increased mutational burden but improved survival in OAC (Noble et al., 2016; Secrier et al., 2016). The infiltration of another immune cell type, M2 macrophages, is promoted by Th2 cytokines. They are characterised by their anti-inflammatory cytokine production resulting in a pro-tumorigenic effect. Increased proportions of these tumour-associated macrophages have been shown to be associated with poor prognosis in squamous oesophageal carcinoma (Sugimura et al., 2015).

A number of clinical trials of immune modulators e.g. PD-1 inhibitors are currently underway in oesophago-gastric cancer. PD-L1 can be expressed by tumour cells, which interacts with PD-1 on effector T cells and suppresses the T cell response. Nivolumab, an anti-PD1 inhibitor, was found in the Phase III ATTRACTION-2 trial of advanced gastric and gastro-oesophageal cancers, in an Asian population, to offer a survival benefit versus placebo (Kang et al., 2017). These have been combined with CTLA-4 inhibitors and have been shown to be superior in combination to a PD-1 inhibitor alone in Phase II studies, but with significant side effects (CheckMAte-032 Study) (Janjigian et al., 2018).

The roles of the immune cells in the progression of Barrett's through the dysplastic stages is less well understood. An increased Th-2 cytokine profile (IL-4, IL-10) and CD4+ T cells in

non-dysplastic Barrett's compared to inflamed squamous oesophagus has been described (Fitzgerald et al., 2002; Moons et al., 2005). However, these profiles were similar in duodenal biopsies (Lind et al., 2012) suggesting that it may be the presence of intestinal metaplasia phenotype causing the shift in immune infiltrate. T regulatory and dendritic cell (DC) densities have been shown to increase in dysplasia but maturation of these DCs may be impaired by factors secreted by the Barrett's mucosa (Somja et al., 2013).

4.4.2 Immune deconvolution

We used GSVA to compute the immune cell compositions of the samples. As with the previous analysis, there were several possible ways to group the biopsies for comparison. The overall grade of the patient, irrespective of the composition of the actual biopsy, or the highest grade captured in the biopsy could be used. As the biopsy grade is harder to be certain of with frozen tissue, we firstly used these enrichment scores to compare the overall patient grades, as we had done above and for the genomic features in Results 1. However, the proportions of immune cell groups by patient grade did not change significantly with progression. So, we then considered the immune compositions based on the highest grade within the biopsy and found significant differences. Firstly, considering the myeloid lineage: we observed a significant decrease from ND to LGD in immature dendritic cells (p value < 0.01) and a trend to a decrease in neutrophils (Figure 49). Both of these cell groups then increased with progression through the dysplasia grades, with a significant rise in neutrophils from LGD to HGD/IMC. Neutrophils are attracted to the TME by chemokines secreted by the tumour cells and they can have both pro- and anti-tumour roles (reviewed in (Galdiero et al., 2013).

A similar pattern was seen in some of the lymphoid lineages: T regulatory cells, B cells and NK CD56dim cells (Figure 50). Cytotoxic T cells did not change significantly. There were no significant increases from HGD to IMC, however the IMC group was small because few samples pathologically contained IMC.

These findings were in keeping with previous pre-sequencing work in BE, which, in particular, have observed a failure of maturation of DCs (Somja et al., 2013), supporting the increasing proportions of immature DCs which we observed.

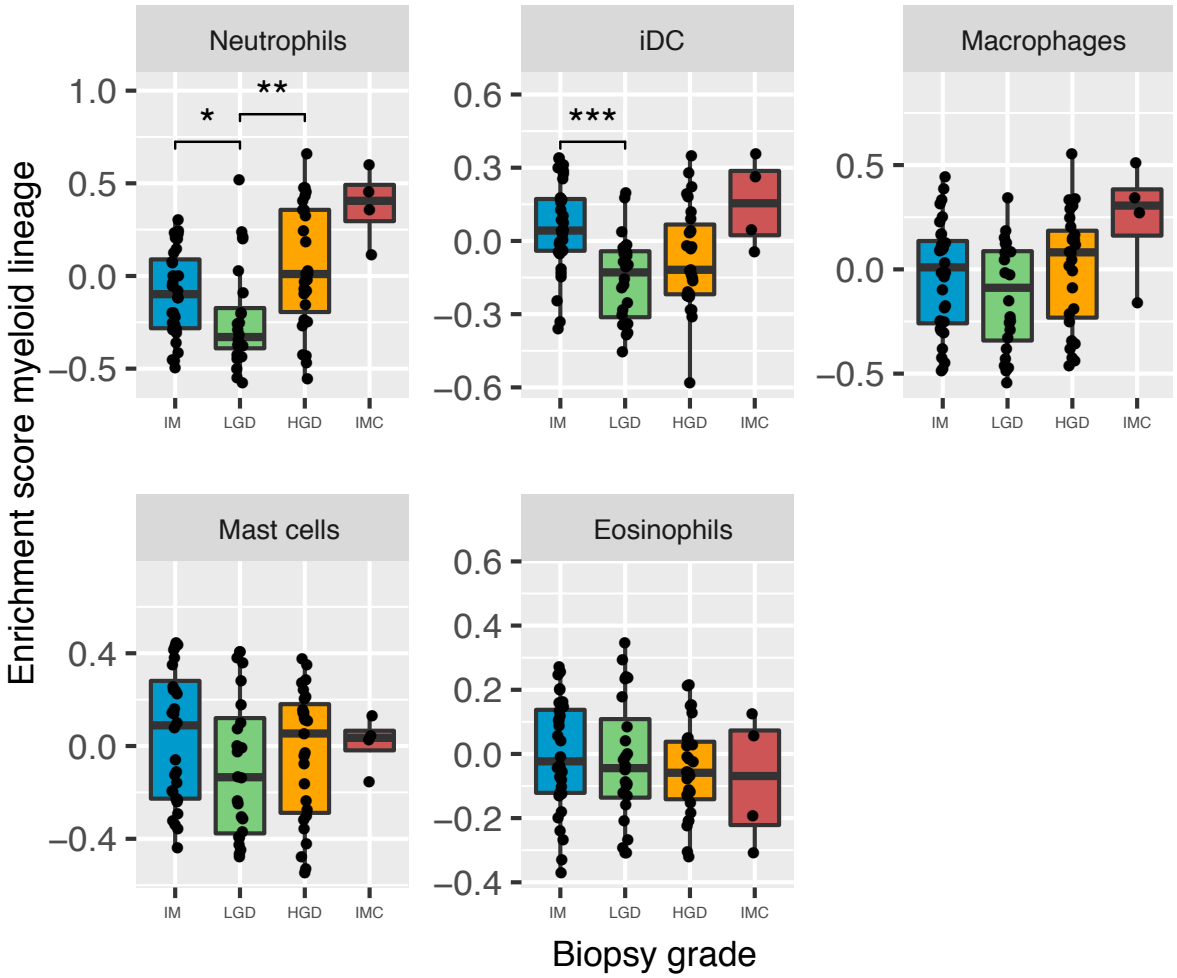


Figure 49 RNA enrichment scores for myeloid immune cell types by biopsy grade

Enrichment scores plotted by biopsy grade for immune cell types in the myeloid lineage. iDC = immature dendritic cells. IM = intestinal metaplasia, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma. The Y axis scale is variable for each cell type. Significant p values are marked with asterisks: * = p value < 0.05, ** = p value < 0.01, ** = p value < 0.001, Kruskal-Wallis test.

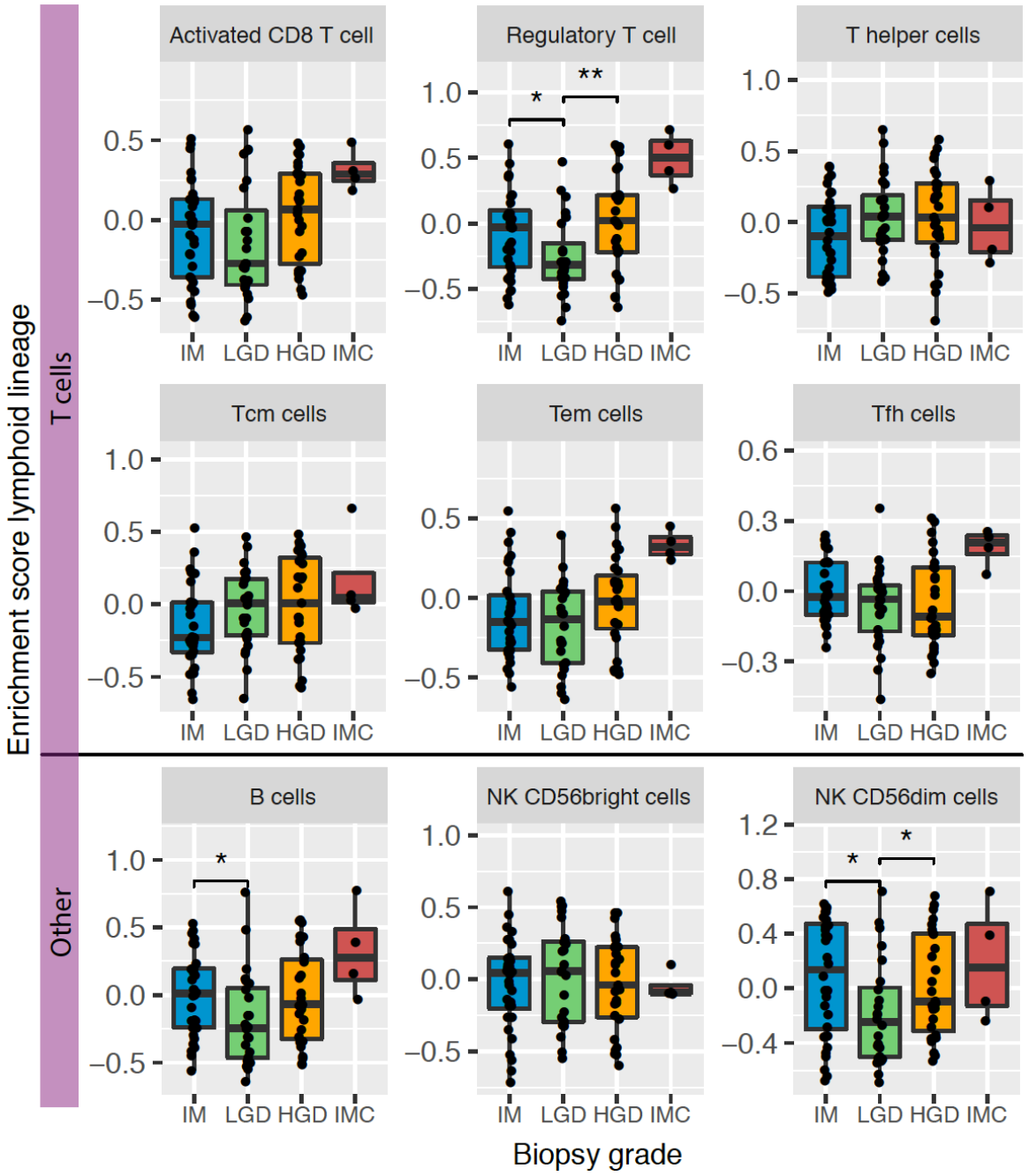


Figure 50 RNA enrichment scores for lymphoid lineage immune cell types by biopsy grade

Enrichment scores plotted by biopsy grade for immune cell types in the lymphoid lineage. Tcm = T central memory, Tem = T effector memory, Tfh = T follicular helper cells. IM = intestinal metaplasia, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma. Y axis scale variable for each cell type. Significant p values marked with asterisks: * = p value < 0.05, ** = p value < 0.01, Kruskal-Wallis test.

Finally, we performed hierarchical clustering using these enrichment scores to see if, overall, there was a change in proportions of cells in progression, when considering all immune cell groups. The grades did not cluster together, but the clustering did highlight two groups within the cohort: those with a higher immune infiltration (across both myeloid and lymphoid lineages) and those with lower (Figure 51). Increased cytotoxic T cell infiltrates are associated with an increased mutational burden and improved survival in OAC (Noble et al., 2016; Secrier et al., 2016) and increasing CNAs have been shown to correlate with reduced expression of markers of CD8+ T cells (Davoli et al., 2017). Using the clustering dendrogram we classified the samples as being either immune high or low and calculated the total mutation burden and % CNA for each sample. However, we did not observe any correlation between immune infiltration and numbers of alterations in this cohort.

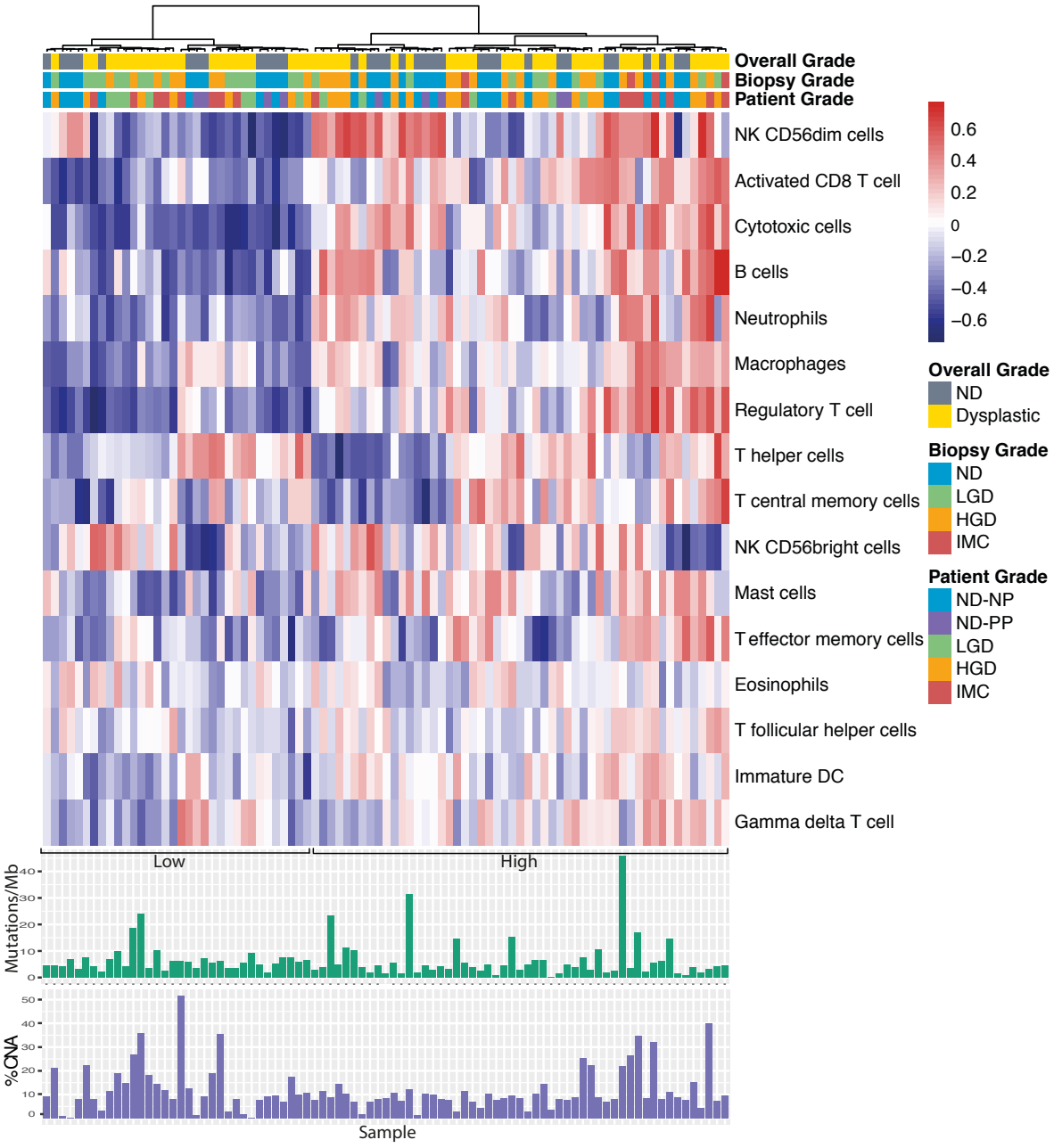


Figure 51 Hierarchical clustering on immune cell type from expression data

Hierarchical clustering of pre-cancer cohort by immune cell enrichment scores, calculated using Gene Set Variation Analysis. Dendrogram splits samples into low and high immune infiltration. Bar charts beneath show mutation burden and copy number aberration percentage for each sample. NK = Natural killer, DC = Dendritic cell. ND = non-dysplastic, NP =non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high-grade dysplasia, IMC = intramucosal carcinoma.

4.5 Summary

We performed bulk RNA sequencing on 125 Barrett's samples and analysed the expression differences seen between non-dysplastic and dysplastic samples. A differential analysis identified 475 deregulated genes. The expression of these genes clustered ND from dysplastic samples well. However, we did not observe individual clustering of the grades of dysplasia.

BE with intestinal metaplasia (IM) shares a number of intestinal phenotypic features with duodenum, but also has functional similarities to the pyloric glands of the stomach. Overall, on consideration of the most variably-expressed genes between these three tissue types, BE split distinctly from both, supporting the incomplete intestinal metaplasia of BE. However, when we considered only the genes that were differentially expressed between ND and dysplastic BE, we found that more than two thirds that were downregulated in progression were highly expressed in duodenum. This included specific genes which are expressed in goblet cells, *TFF3* and *MUC2*, which have previously been shown to be good biomarkers of BE (McIntire et al., 2011; Ross-Innes et al., 2015b). Histologically, we observe a reduction in IM in dysplastic tissue (Naini et al., 2016; Odze, 2006) and these results support this observation: with progression to dysplasia there is a downregulation of expression of genes that convey the IM phenotype. Clinically, TFF3 staining sensitivity has been shown not to be reduced in the higher grades when used as a biomarker (Ross-Innes et al., 2015b). So, this loss of IM phenotype is an important finding biologically but would not necessarily affect BE biomarker design because the genes are still expressed in dysplasia, just at a lower level.

A previous microarray study considered the 'intestinal-like signature' that they saw in BE and found it to persist in OAC (Duggan et al., 2016). However, very small numbers of normal tissues were used to form the signature (3 squamous oesophagus, 3 colon, 3 duodenum). We did not see this preservation in dysplasia however, it would be interesting to compare the expression of our gene signature in OAC too.

Pathway analysis of the genes downregulated from ND to dysplastic identified *HNF4A* as a regulator effect network. *HNF4A* encodes a transcription factor that is usually found in primitive intestinal cells in embryonic development. It has recently been shown to induce chromatin opening in normal oesophagus cells resulting in a Barrett's-like chromatin signature (Rogerson et al., 2019). The downregulation of this network in dysplasia would fit

with the loss of the IM phenotype. A previous recent whole transcriptome study had findings that, on first look, do not seem in keeping with our results. They found *HNF4A* to be upregulated in OAC and expressed (but at a lower level) in NDBE (n=14) and LGD (n=8) (Maag et al., 2017). However, they mainly compared OAC (n=12) to normal squamous oesophagus (n=17), rather than GC or D2. It fits that all the glandular tissues had higher expression of *HNF4A* than squamous tissue. In our analysis, although comparisons were between the dysplasia and ND, we did not see upregulation of *HNF4A* itself. A comparison of our data to OAC may help to resolve this.

We observed a much higher number of significantly upregulated genes with progression (352) than those downregulated (123). In particular, there was a significant increase in expression of a number of genes downstream of the ERK/MAPK pathway. This pathway has previously been shown to be activated in BE cell lines by acid exposure (Jaiswal et al., 2006; Morgan et al., 2004; Souza et al., 2002). Several members of the cascade e.g. *EGFR, MYC* and *KRAS* and have been shown to be amplified in OAC, both in historic studies (e.g.(Sommerer et al., 2004)) and recent genomic studies (Frankell et al., 2019). However, in Results 1 we only observed *MYC* to be amplified in one dysplastic sample. AP1 (FOS/JUN), a transcription factor also downstream in the ERK/MAPK pathway, has been implicated as a key regulator of expression in OAC (Britton 2017). Most studies of these pathways have been on cell lines or microarrays on small numbers of human samples. Paterson et al. compared gene expression in NDBE, dysplasia and OAC using microarrays but found the MAPK pathway enrichment in NDBE and dysplasia to be the same, and only enrich further in OAC, when using a small number of genes (Paterson et al., 2013). Our results indicate that the upregulation of the pathway is important in the transition from a ND to dysplastic phenotype, which has not previously been shown. The important next step will be to look at the methylation of the genes in this pathway as this is likely to explain the changes in expression. If recurrently hypomethylated promotors can be identified, then it may be possible to derive a biomarker from this. It would also be necessary to perform functional experiments to prove this with cell lines.

In the immune analysis of the expression data, we found that the immune composition depended more on the specific grade of dysplasia surrounding the immune cells rather than the overall patient grade. This is different to the genomic analysis, where a field effect seemed to occur. The three immune cell types which best correlated with progression through the grades of dysplasia were neutrophils, T regulatory cells and immature dendritic

cells. All of which have been shown to have roles in tumourigenesis (reviewed in (Togashi et al., 2019; Tran Janco et al., 2015)).

In the differential analysis we showed *CXCL1* expression to be significantly higher in dysplasia, which may explain the observed increase in neutrophils. Hepatocellular carcinoma cells have been shown to secrete chemokines which attract neutrophils. The neutrophils, themselves, secrete pro-angiogenic proteins which promoted progression (Kuang et al., 2011). A number of other studies have also shown neutrophils to have pro-tumourigenic properties (reviewed in (Tecchio et al., 2013)) but this has not been studied in the progression of BE.

For all three immune cell types (neutrophils, T regulatory cells and immature dendritic cells) we observed an initial reduction from ND to LGD, with significance reached ($p < 0.001$) for iDCs. These findings need to be validated in larger cohorts by IHC. We attempted a new multiplex IHC technique which allows the simultaneous analysis of multiple immune types from one tissue cut. However, it was unsuccessful in this cohort because the small biopsy cuts lost adherence to the slide and washed away. Instead, we need to focus on the 3 immune cell types that had significant changes and perform traditional IHC separately for each of them.

It would also be interesting to look at larger resection samples composed of multiple grades and compare the immune composition to grade within a sample. Furthermore, the iDC levels need to be compared to mature DCs and, whilst there was no overall change in the numbers of macrophages, the ratio of M1/M2 macrophages has been shown to be important in cancer progression and this should also be looked at in future work.

In the genomic analysis, the samples followed a gradual continuum with progression. In the expression analysis we observed a similar gradual progression of increasing expression when performing a principle component analysis. However, when we grouped genes to analyse specific pathways, we observed more of a distinction between the expression in non-dysplastic versus dysplastic samples. Overall, it would seem that the continuum of SNVs, CNAs and SVs leads to a continuum of expression changes. We hypothesise that when this altered expression is for genes that fall into the same pathway, the balance is tipped, the pathway is upregulated, and progression occurs.

5. Results 3: Clonal heterogeneity in Barrett's oesophagus

Aim 2: Consider the heterogeneity and clonal evolution of BE segments and how this may influence progression.

- Examine the genomic heterogeneity within Barrett's oesophagus segments.
- Use clonal and subclonal mutations and copy number alterations to understand the clonal evolution of Barrett's oesophagus.

5.1 Introduction

In cancer, the importance of intra-tumoural heterogeneity (ITH) as well as heterogeneity between tumours is becoming increasingly recognised. If different clones within the tumour respond differently to therapy then the degree of ITH may affect treatment choice and can determine therapeutic response and survival e.g. (Jamal-Hanjani et al., 2017; Murugaesu et al., 2015). ITH has also been shown to be important in pre-invasive disease. Maley et al. calculated a diversity score in BE using: aneuploidy and tetraploidy, by flow cytometry; loss of heterozygosity of the loci containing *CDKN2A* and *TP53* using fluorescent in-situ hybridisation (FISH); and *CDKN2A* and *TP53* mutation (Maley et al., 2006). They showed that increased clonal diversity at baseline could predict the risk of progression to cancer. More recently, a single cell analysis using a panel of FISH probes also found the baseline diversity to predict progression (Martinez et al., 2018). Next generation sequencing has not widely been used to study the diversity within BE, but this approach would enable a more extensive analysis without a priori knowledge or assumptions about which loci are likely to be informative. Multilevel WES has previously been performed on five patients with Barrett's adjacent to cancer to look at the clonal relationship of the cancer to BE (Stachler et al., 2015). A lot of heterogeneity was observed, in that not all BE biopsies from a given patient were clonally related to the tumour (as we observed in our analysis in Results 1: 3.3.3). Two of the five patients had missense *TP53* mutations in all their biopsies. This clonality led them to speculate that *TP53* mutation may occur prior to dysplasia: earlier than previously thought. They also observed oncogene amplification confined to individual biopsies, suggesting that it may be a later phenomenon. In a previous study from our laboratory, WGS with further targeted amplicon sequencing was used for an in-depth analysis of a single patient with IMC (Ross-Innes et al., 2015a). This study found evidence for an initial clonal sweep with dysplasia developing from multiple different clones.

Together, these studies confirm the heterogeneity of BE, and suggest that a single biopsy from a BE segment does not represent the whole lesion. To date, studies have focussed on either a few specific events or on BE adjacent to cancer. The dynamics of the complete genomic landscape across a segment have not been studied in detail. We also do not know whether the structural variation follows the same patterns of heterogeneity as the SNVs and CNAs. In our WGS, we focussed on using the highest grade possible within the segment, with the best cellularity in order to compare grades between patients. Here we aimed to

elucidate how these biopsies related to their surrounding BE, and how grade and distance affected this.

5.2 Multilevel cohort selection

In order to study the intralesional heterogeneity in our cohort, multiple levels were sequenced from a subset of patients in my overall BE cohorts. Specifically, 15 additional levels from 6 cases were included, for which further good quality frozen biopsies were available (containing a good percentage of glands and no squamous, but irrespective of grade). Figure 52 details these cases. The overall grade of the patient and highest grade seen within each individual biopsy are given. All biopsies were taken at the same time as, or within 6 months of, the previously-sequenced biopsies as it was felt that there would be unlikely to be any genomic changes in such a short period. This gave us 3-4 levels per case.

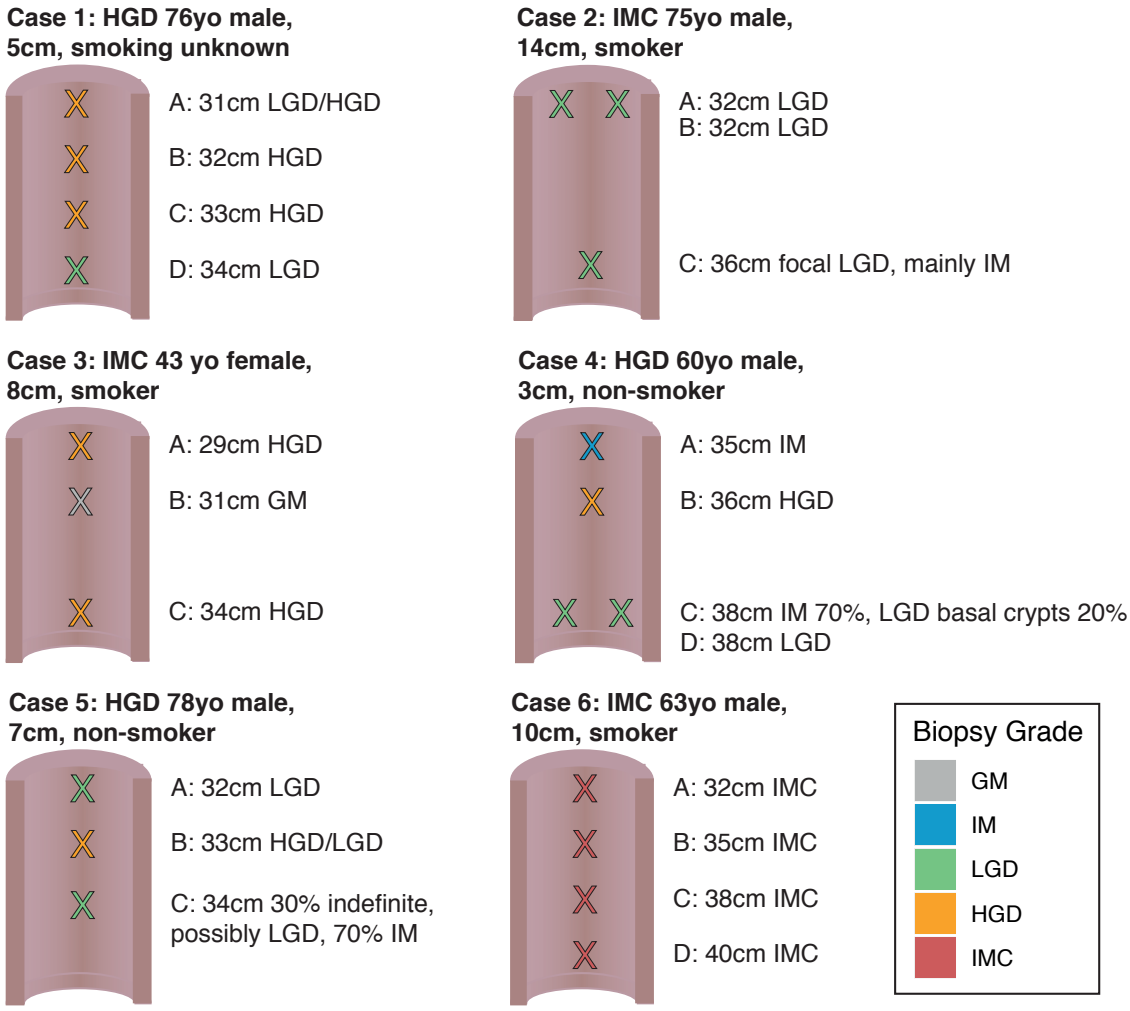


Figure 52 Whole genome sequencing of multilevel cases

Frozen biopsies taken from multiple levels of 6 cases identified for whole genome sequencing to give 21 levels in total. For each patient the overall grade at that timepoint, age, sex, maximum length of BE and smoking status is given in the heading. Images depict levels (distance from the incisors in cm) at which biopsies were taken and the pathological grade of the biopsy. Indefinite for dysplasia means that the biopsy was difficult to grade. GM = gastric metaplasia, IM = intestinal metaplasia, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma, yo = years old.

5.3 Genomic features of the multilevel cases

Firstly, the total mutation burden (TMB), copy number aberration (CNA) and numbers of structural variants (SVs) between biopsies taken from single individuals were compared. Figure 53 shows this for each of the 6 cases. Each case presented a very different profile. The TMB ranged from 0.24-15.1 mutations/Mb; median 7.0. This variation was irrespective of the grade composition of the biopsy. In cases 1,4 and 6, all the biopsies from a given individual had similar profiles. In contrast, case 2, 3 and 5 showed more variation across biopsies. This diversity was also very noticeable when looking at the CN profiles of individual cases (Figure 54). For example, whilst the biopsies in case 6 have similar total % of genomes with CNAs, the profiles are heterogeneous.

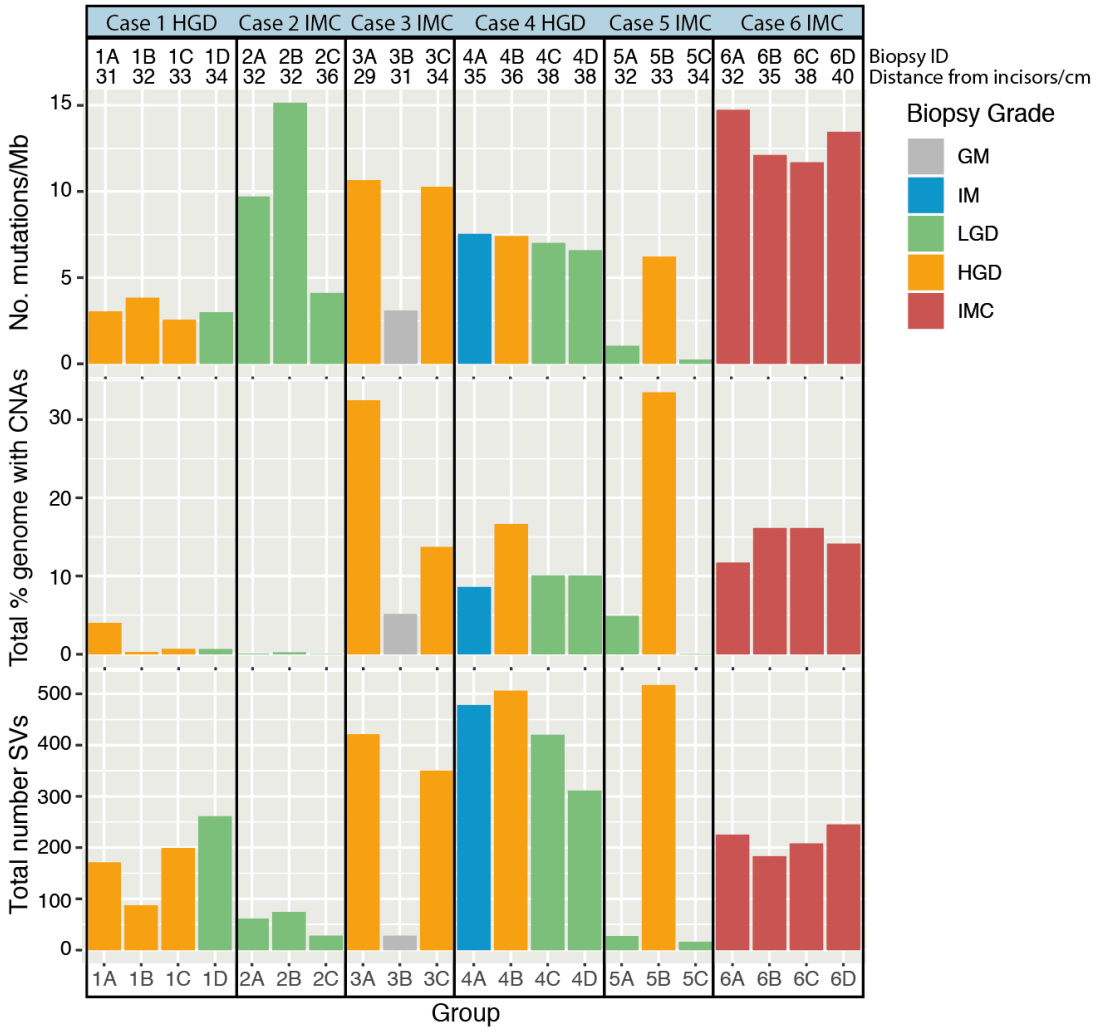


Figure 53 Genomic features of the multilevel cohort

Biopsies from individual cases are grouped, with the overall grade and the levels from which the biopsies taken in cm. The colour of each bar represents the highest grade of BE within that biopsy. The 3 bar charts give measurements mutation burden, total % of genome altered by CN and total number of structural variants. GM = gastric metaplasia, IM = intestinal metaplasia, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma, CN = copy number, SV = structural variant.

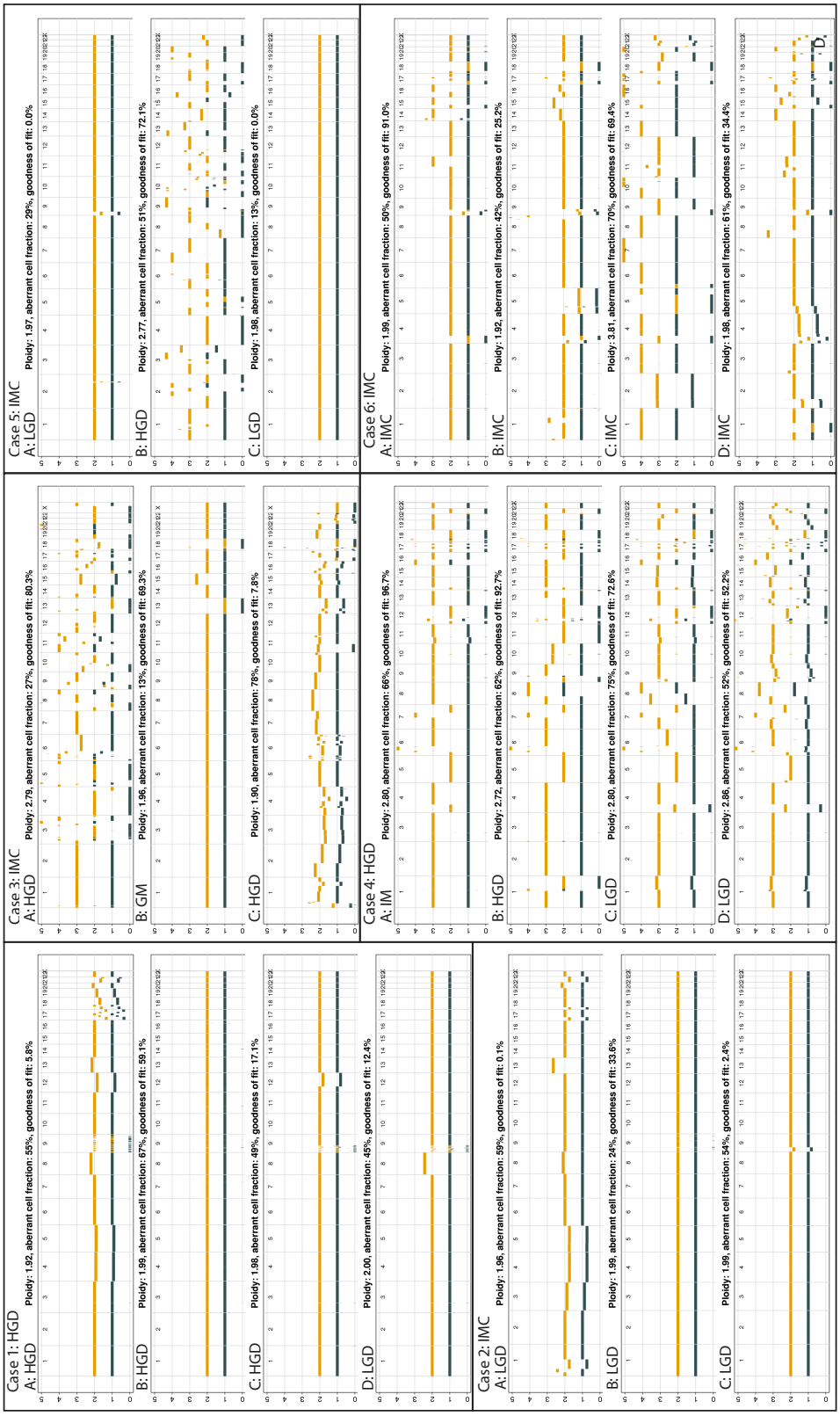


Figure 54 Genome wide copy number profiles of multilevel cases

Genomic copy number profiles from the Battenberg output for each biopsy. Case number and overall grade of the patient are indicated above each set of plots. Biopsy ID and grade are detailed above individual plots. Chromosome number is listed along the plot from 1 to X. Yellow lines represent the total copy number of each region, grey lines represent the B allele frequency. GM = gastric metaplasia, IM = intestinal metaplasia, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

We next considered the proportions of each mutational signature contributing to each biopsy and found that, on the whole, there was preservation of the proportions mutational signatures between biopsies from a case (Figure 55). Where heterogeneity did occur e.g. case 1, the biopsies that were different did not match those that stood out as different with the genomic features above (mutation burden, total CNA, SV count).

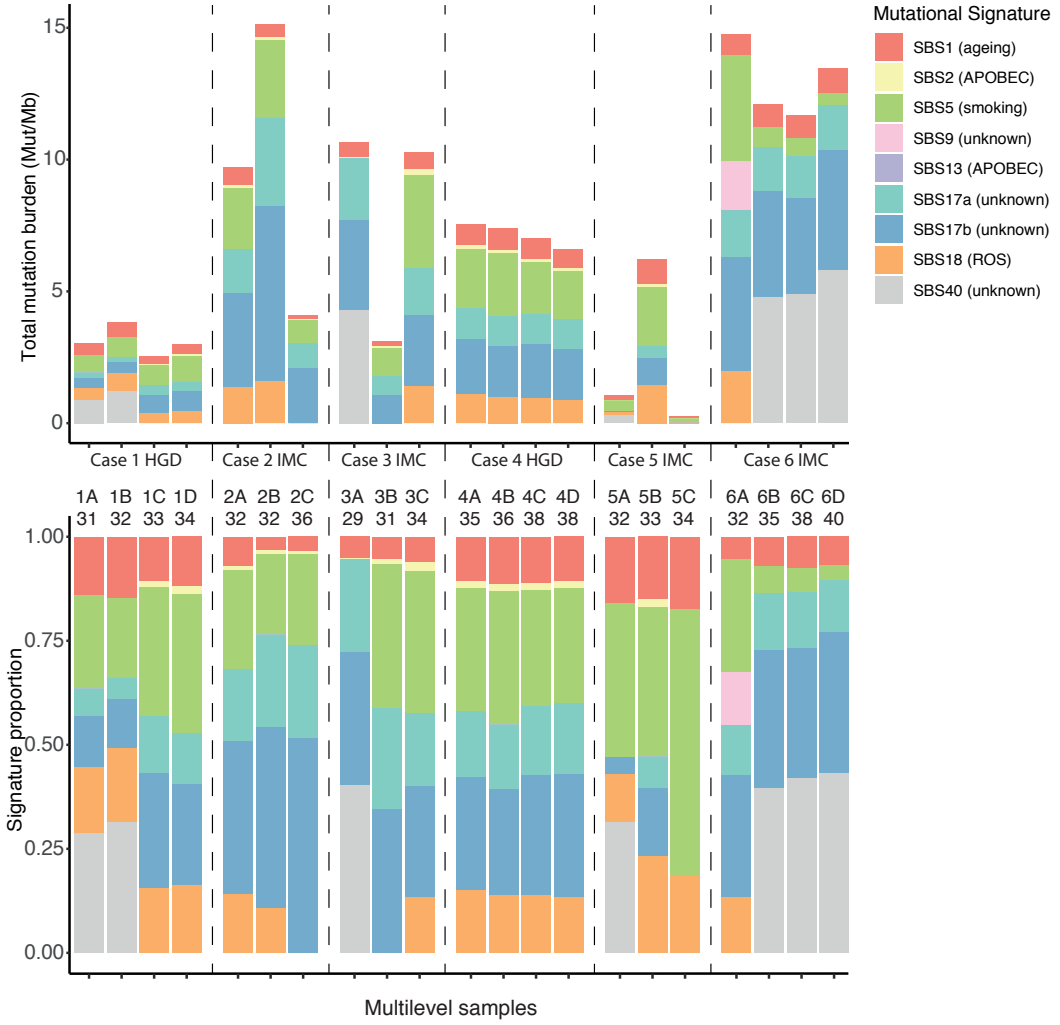


Figure 55 Mutational signatures in the multilevel cases

The proportions of each mutational signature contributing to biopsies relative to the total mutation burden. The upper plot shows proportions of each signature within the total mutation burden. The lower plot shows the overall proportions of each signature. Known aetiologies of the signatures are given in parentheses (https://cancer.sanger.ac.uk/cosmic/signatures). HGD = high grade dysplasia, IMC = intramucosal carcinoma, SBS = single base substitution.

The SV signature profiles were also generally preserved within cases (Figure 56). However, where there was variation, this too was not concordant with the variability of the genomic features. Biopsies in case 4, which had similar genomic profiles, also had similar SV signature proportions and mutational signature proportions. However, in case 1, the four biopsies had similar genomic profiles, but biopsy B had a different SV signature pattern, with signatures 1 and 2 present but an absence of SV signatures 3 and 4. Furthermore, two different mutational signature profiles were seen in first two versus the second two biopsies. Other cases had similar SV signature proportions despite the genome-wide CN profiles above looking very different e.g. cases 2 and 3.

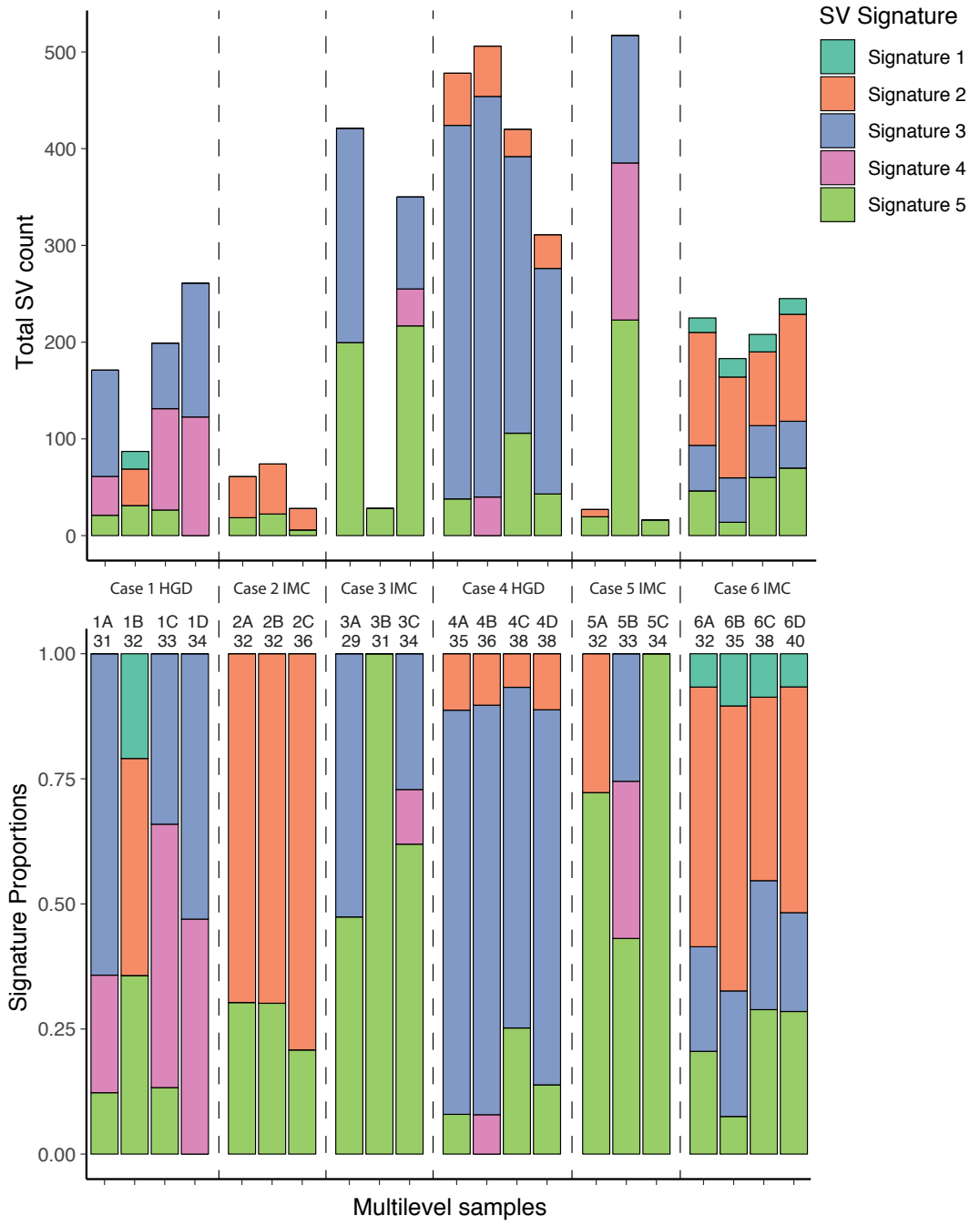


Figure 56 Structural variant signatures in the multilevel cases

The proportions of each SV signature contributing to biopsies relative to total number of SVs. Upper plot shows proportions of each signature within the total SV count. Lower plot shows the overall proportions of each signature. HGD = high grade dysplasia, IMC = intramucosal carcinoma.

Lastly, we considered the SNV and CN driver gene mutation events and how the biopsies compared (Figure 57). Driver events were identified in 19/21 biopsies. No driver event was identified in biopsy 1C, despite containing IMC. As above, the heterogeneity was marked, but patterns could be seen when comparing the driver gene profiles to the CN alterations shown in Figure 54. For example, biopsies 1A and 5B are *TP53* mutant and have a very different genomic CN profile to their neighbouring wildtype biopsies.

Focussing on *TP53* mutations specifically, in 2 of the 6 cases missense mutations in *TP53* were seen in all the biopsies (cases 3 and 4). In case 4, all biopsies had the same position *TP53* SNV: p.R273C a *TP53* hotspot. However, in case 3 only 2 of the biopsies shared the same *TP53* mutation position. This suggests that *TP53* mutation was an early event in these cases, and before WGD as this was only observed in 3A.

Case 5 only has a *TP53* mutation in biopsy B: so, in this case *TP53* mutation must have occurred later. This biopsy was also the only one to have undergone WGD. Interestingly, Biopsy 3B was described pathologically as gastric metaplasia, yet it had a *TP53* mutation. Gastric metaplasia alone is considered indolent with almost no risk of progression. *TP53* was also mutated in the IM biopsy 4A. Case 6, was clinically unusual because on endoscopy there appeared to be 4 distinct areas of IMC. In terms of SNV, CNV and SV numbers, there was little difference between the 4 biopsies. The genomic CN profile highlighted somewhat more variation. However, the heterogeneity became most clear when considering the driver events: with GATA6, *PTEN*, *ERBB2* and *MYC* mutations seen in different biopsies. The *TP53* and *ERBB2* mutations were the same positions in the 3 biopsies they were in.

WGD was called in 7/21 biopsies. In three patients it was only in one of the biopsies, suggesting that it may have been a late event. However, in case 4, all four biopsies exhibited WGD, indicating that it may have occurred earlier and led to the expansion of a single clone.

Overall, the timing of *TP53* mutation in these cases was variable: an early event in some cases and late in others. However, WGD was generally a later event and only occurred in the *TP53* mutant biopsies. This fits with our WGD analysis in Results 1. However, the pathway to cancer appears to be more complicated than the model suggested by Stachler et al. (Stachler et al., 2015). In their analysis of BE adjacent to cancer, they observed early *TP53* mutation and WGD. We observe there to be more variation than this, in particular the observation of *TP53* mutation and WGD in only one of the dysplastic biopsies within a segment.

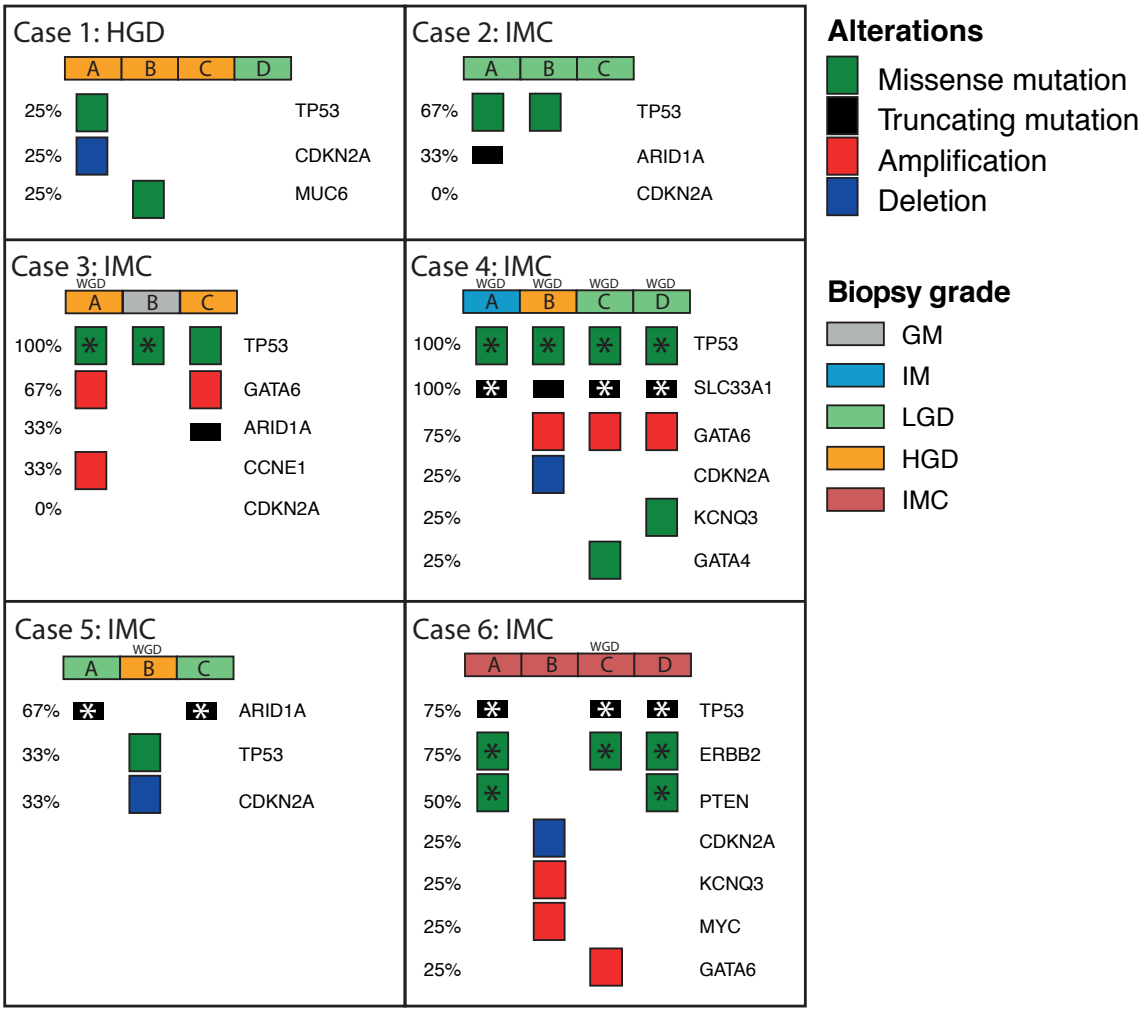


Figure 57 Driver mutations in the multilevel cases

Single nucleotide variants, amplifications and deletions in known driver genes per biopsy. Asterisks identify mutations in same genomic positions across biopsies from a single case. Patient grade in headings, biopsy grades colour coded, with lettering referring to the biopsy ID. Only driver genes with mutations are listed. Driver gene list as used in Results 2. Any excluded genes are not mutated. Whole genome doubling (WGD) noted above biopsies. GM = gastric metaplasia, IM = intestinal metaplasia, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

5.4 Clonality analysis

The next aim was to analyse the clonality of the cases in order to construct phylogenetic trees. To do this, all pre-filtered calls for SNVs and their positions were analysed to see if the mutations were shared between biopsies. The overall number of SNVs called, and the degree of overlap, varied vastly from case to case (Figure 58). The four dysplastic biopsies in case 1 (HGD) had similar numbers of SNVs and % of genome altered by CNAs between them in the above analysis, yet only 1.2% of the mutations were shared between all the four biopsies (median number SNVs 9,503; range 9,362-11,636). Conversely, case 4, also HGD, had 4 biopsies ranging from IM to HGD. The median number of SNVs per biopsy was higher at 21,514 (range 20,566-22,200) and 81.7% of this median (17,578 SNVs) were seen in all four cases.

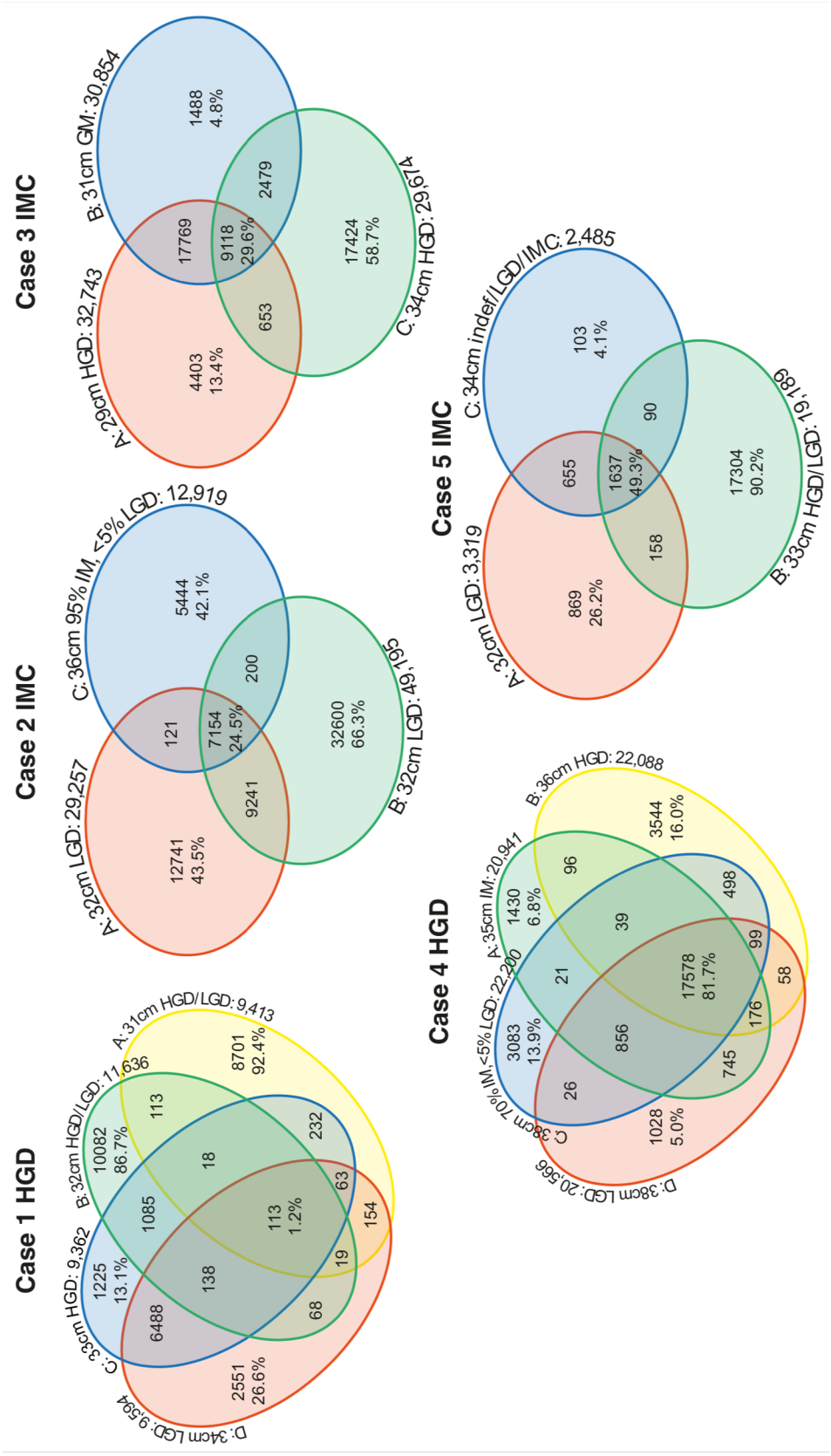


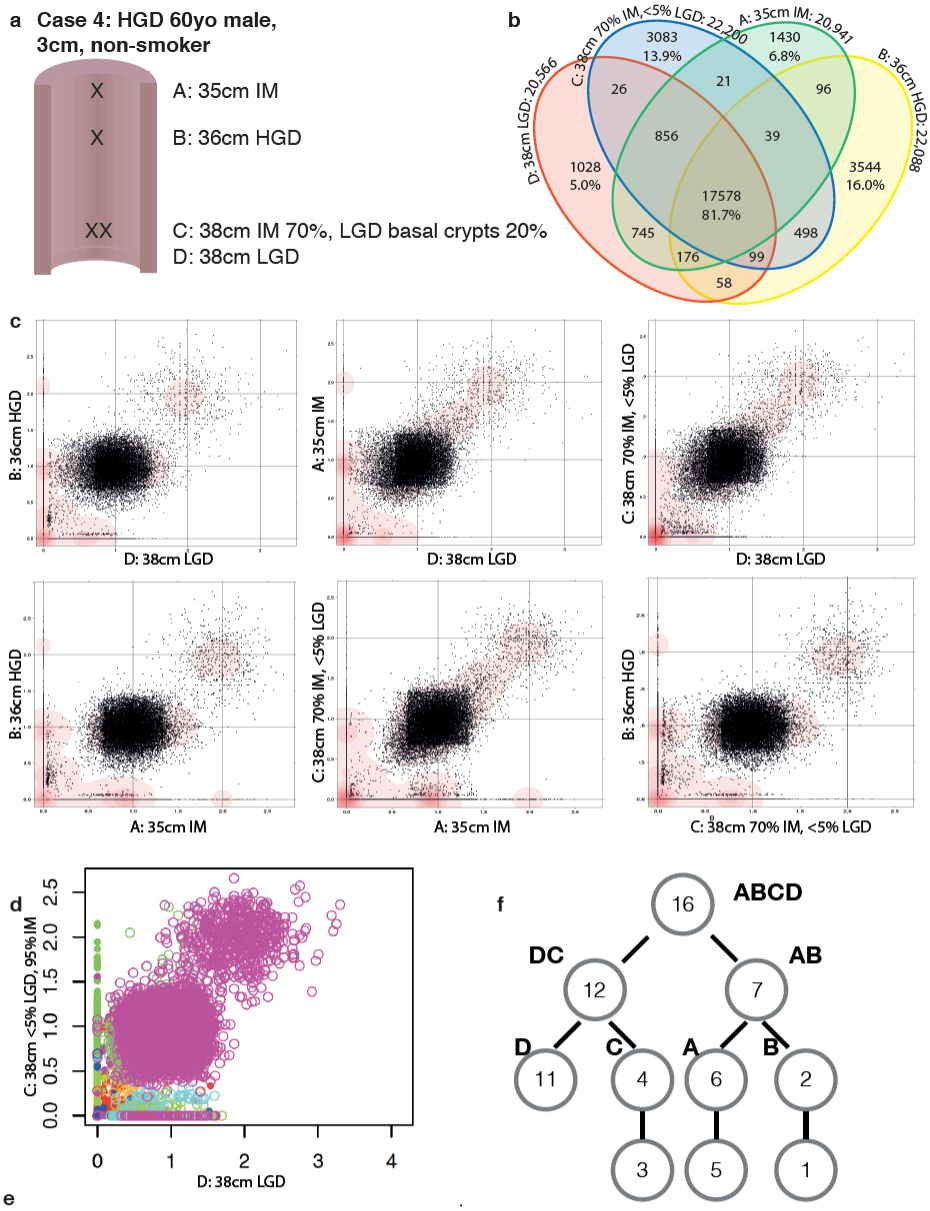
Figure 58 Venn diagrams of SNV overlap for each multilevel case

Total mutation numbers of all SNVs in all positions for each biopsy, with overlap indicating shared mutations between biopsies. Patient grades are given in the headings. Individual biopsy grades and level on each ellipse. Percentage of mutations unique to each biopsy given. Percentage shared between all biopsies from one case calculated using the median mutation number of the biopsies in the case. GM = gastric metaplasia, IM = intestinal metaplasia, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

DPClust v2.2.5 (Nik-Zainal et al., 2012a) was used to model clonal expansions by calculating the cancer cell fraction (CCF) of each mutation. This is the proportion of tumour cells (in this case BE cells) in which the variant is present. Estimating the CCF depends on the local CN and the cellularity (proportion of BE cells in the biopsy). A variant with a CCF of 1 means that it is in all the 'cancer' cells and is considered clonal, whereas any variants with CCFs < 1 are subclonal.

Figure 59c shows scatterplots of pairs of biopsies from Case 4. For each SNV the CCF in one biopsy is plotted against the CCF in the other biopsy. The clonal cluster can be seen at the coordinates x =1, y=1. Case 4 has a clear clonal cluster shared by each biopsy, comprising 18,969 mutations (pink cluster 16 in Figure 59d). This gave rise to 2 divergent subclones: 7 (390 mutations) and 12 (628 mutations) (Figure 59e, f). These each gave rise to two further subclones, resulting in unique subclones within each of the four biopsies. Clusters with fewer than 1% of the total number of mutations were excluded.

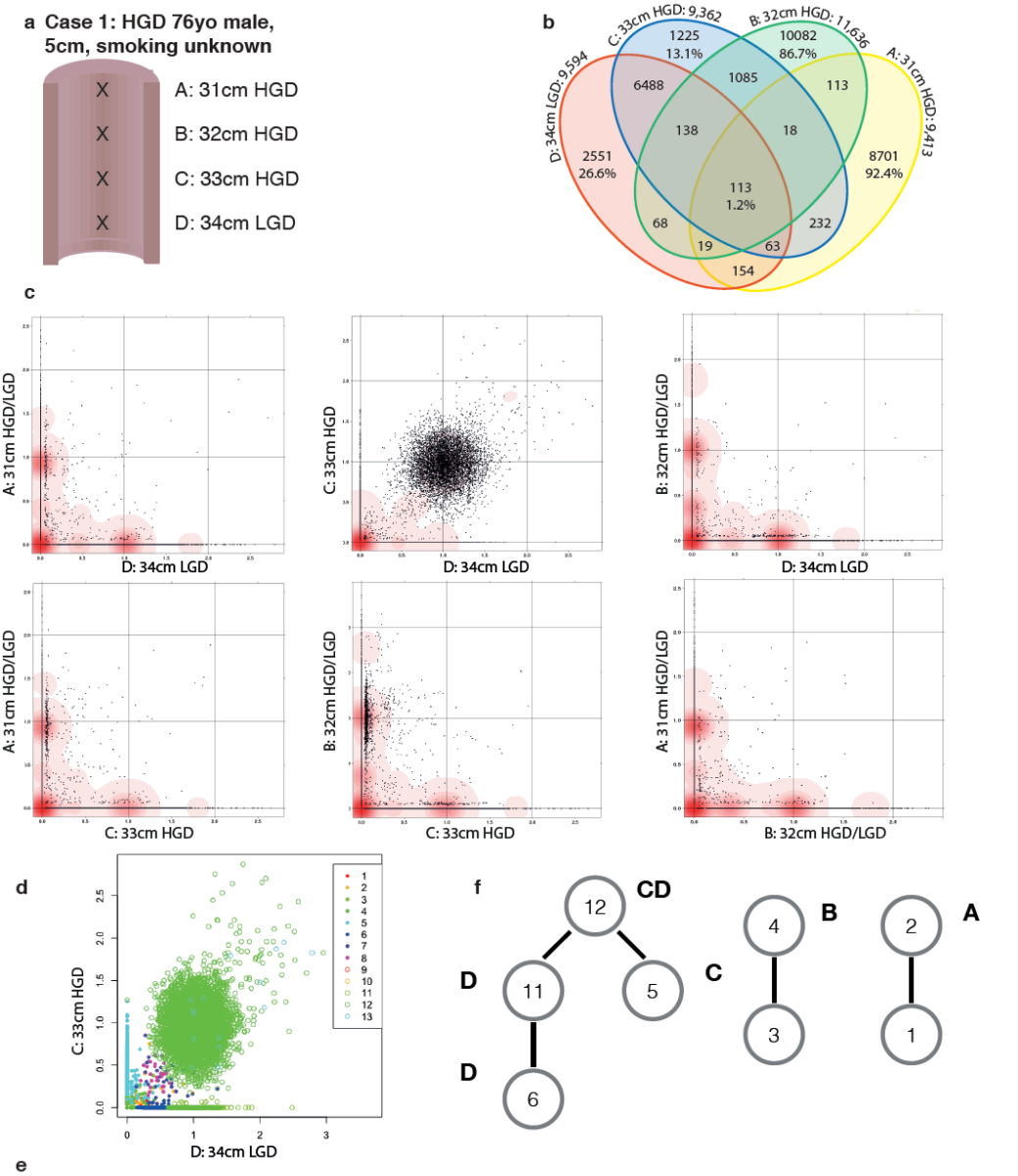
Case 1, in contrast, had minimal sharing of variants and no common ancestral clone. Subclone 12 (6768 mutations) was present with a CCF of 1 in biopsies C and D (Figure 60c). Subclone 2 was seen in biopsy A and subclone 4 in biopsy B. This meant that there was no common clone from which all BE cells arose, which in this case, does not fit with the idea of there being an initial clonal expansion.



Cluster	Cancer cell fraction				No. mutations
	D: 38cm LGD	A: 35cm IM	C: 38cm <5% LGD	B: 36cm HGD	
1	0	0	0	0.338	1822
2	0	0	0	0.941	2095
3	0	0	0.357	0	660
4	0	0	0.825	0	790
5	0	0.319	0	0	230
6	0	0.938	0	0	2956
7	0	0.967	0	0.943	390
8	0.368	0.946	0.805	0.978	68
9	0.388	1.229	0.270	0.413	6
10	0.474	0.482	0.514	0.507	58
11	0.540	0	0	0	999
12	0.611	0	0.868	0	628
13	0.639	1.004	0	0.941	113
14	0.807	0	0.861	0.861	40
15	0.828	0.948	0.971	0	9
16	1.001	0.991	0.989	0.989	18969

Figure 59 Case 4 clonality analysis: all subclones arising from one common ancestral clone

a. Demographics, biopsy levels and pathology. **b.** Venn diagram of overlap of SNVs. **c.** DPClust output showing paired biopsy comparisons of cancer cell fractions (CCF) of each SNV. CCF on the x and y axis, with clonal cluster at 1,1. **d.** DPClust assignment of cluster to SNVs based on CCF. **e.** Cluster output and number of mutations in each cluster. CCF for each biopsy. Clusters with <1% of the total number of mutations are filtered out. **f.** Phylogenetic tree construction based on the cluster CCF per biopsy in e. IM = intestinal metaplasia, LGD = low grade dysplasia, HGD = high grade dysplasia.



e

Cluster	Cancer cell fraction				No. mutations
	D: 34cm LGD	C: 33cm HGD	B: 32cm HGD	A: 31cm LGD/HGD	
1	0	0	0	0.406	1593
2	0	0	0	0.954	7596
3	0	0	0.367	0	4668
4	0	0	1.027	0	6741
5	0	0.338	0	0	1281
6	0.456	0	0	0	2220
7	0.521	0.387	0.311	0.351	33
8	0.507	0.428	0	0	27
9	0.792	0.792	0	0.818	2
10	0.955	1.053	0.615	0	4
11	0.999	0	0	0	228
12	1.026	1.006	0	0	6768
13	1.274	0.946	0.607	0.941	15

Figure 60 Case 1 clonality analysis: distinct, unrelated clones

a. Demographics, biopsy levels and pathology. **b.** Venn diagram of overlap of SNVs. **c.** DPClust output showing paired biopsy comparisons of cancer cell fractions (CCF) of each SNV. CCF on the x and y axis, with clonal cluster at 1,1. **d.** DPClust assignment of cluster to SNVs based on CCF. **e.** Cluster output and number of mutations in each cluster. CCF for each biopsy. Clusters with <1% of the total number of mutations are filtered out. **f.** Phylogenetic tree construction based on the cluster CCF per biopsy in e. IM = intestinal metaplasia, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

Figure 61 shows the trees constructed for the other three cases analysed for clonality. All biopsies in cases 2 and 3 arise from a common clone. Whereas case 5 is complicated: clone 2 (17,567 mutations) is clonal in biopsy B, but not present in A or C. However, clone 4 (3061) is clonal in A and C and subclonal in B. This could possibly be explained by clones converging.

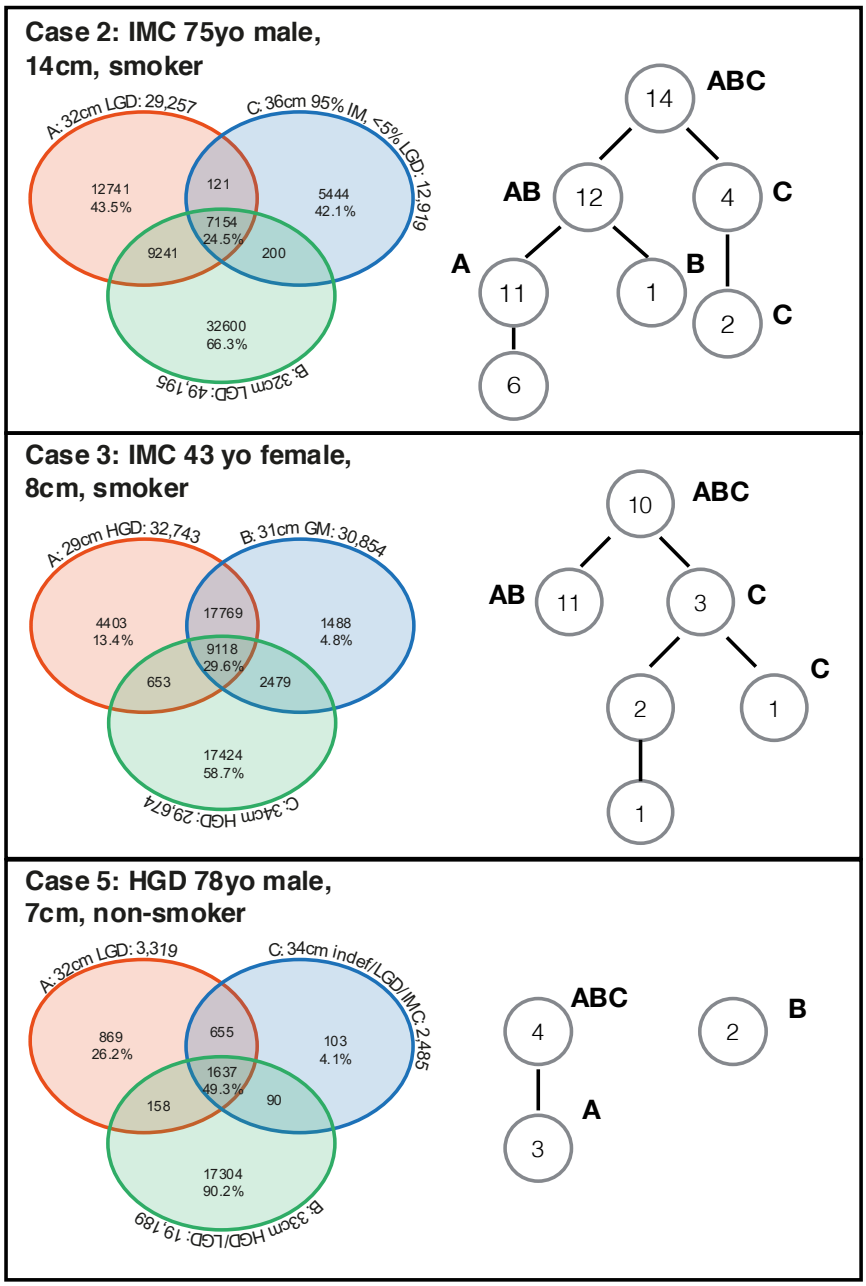


Figure 61 Phylogenetic trees for multilevel cases.

Trees constructed using the cancer cell fraction of clusters of mutations defined by DPClust. Patient demographics and grade given in headings, with age, sex, maximum Barrett's length and smoking status. Venn diagrams as in Figure 58. Phylogenetic tree construction based on the cluster CCF per biopsy.

5.5 Summary

We analysed the intralesional heterogeneity of 6 dysplastic BE segments by performing WGS on multiple levels from each case. It has previously been shown with using more targeted variant analyses that BE segments are likely to be polyclonal and have a high degree of intralesional heterogeneity. However, this has not been evaluated on a cohort of patients using WGS. WGS is perhaps more suited to this purpose than WES or mutation panels because intronic passenger mutations are key in clonal and subclonal analyses. This is likely to be particularly true in BE, which is highly mutated but, unlike most cancers, carries fewer clonal mutations.

The 6 multilevel cases highlighted just how heterogenous BE is, with each case needing to be considered completely independently. However, within cases, the total mutation burden, total % of genome with CNAs and total numbers of SVs were surprisingly constant between biopsies, irrespective of grade. This is reassuring because in earlier analyses we used the overall patient grade over the grade of the actual biopsy. On the whole, mutational and SV signature proportions were preserved between biopsies from a case, supporting their early formation. More heterogeneity was observed when considering the driver gene events, WGD and patterns of genome-wide CNAs, indicating these to be later events. This may suggest that the overall burdens of the different genomic aberrations are determined by the local environment and exposure to specific mutagens, and so there is a consistency between biopsies. But the specific genes disrupted are stochastic. This leads to the development of multiple clones which are heterogeneous mainly because of their spectrum of driver gene alterations i.e. driver gene alterations drive the heterogeneity observed within a BE segment.

Five of the cases underwent a clonality analysis. Only 3 of the 5 cases had a common ancestral clone, from which all subclones were derived. This may imply that BE does not always originate with an initial clonal sweep, or, after the sweep, completely new, independent clones can arise within the segment. We observed clear heterogeneity between biopsies for CN, driver gene mutations and WGD. In order to understand more about the clonal evolutionary paths in the progression of BE, future work needs to include multilevel non-dysplastic cases to compare the heterogeneity seen and include both non-progressor and pre-progressor cases. Furthermore, the construction of formal phylogenetic trees, with the number of mutations represented by the length of the branch, and driver mutation annotation, will be informative regarding the time between clonal sweeps and the key mutations

involved. A divergence score, adapted from those previously published, needs to then be applied. Divergence could then be correlated with progression.

We were surprised to see such a high mutation burden in the gastric metaplasia (GM) biopsy of case 3, when considering all of the unfiltered calls (30,854). In the genomic analysis, we had used the filtered Strelka variant calls, which had called 9681 SNVs. We looked at the variant allele frequencies (VAFs) of this sample which showed that 14,202 of the variants had a VAF < 0.05, and so had not been called by Strelka. This may be because a GM biopsy is of lower cellularity and would be similar to sequencing normal stomach: there would unlikely be many high VAF mutations. These low VAF mutations overlapped with another biopsy, raising the possibility that Strelka may be missing some of these. However, in calling these, a lot of background noise would likely also be called. We used stringent filters in our genomic analysis in order to be highly confident of our calls. We do not have further GM samples to compare to because we avoided sequencing them. Overall, we did not see this problem of low VAF calls with other samples: for the other two biopsies in case 3, one had 417 mutations with a VAF < 0.05 and the other had 702.

Finally, one key advantage of WGS is that the structural variation within a biopsy can be analysed. We performed only a superficial SV analysis, considering count and SV signatures. For future work, an in-depth comparison of the SV types, and specific genes affected between clones would give further information about the timing of these events. DPClust does not take SVs into account. SVClone is a computational method for inferring the cancer cell fraction of structural variant breakpoints from whole-genome sequencing data (Cmero et al.). It would be interesting to use this tool and compare it to the clonality estimates made using SNVs and CN.

6. Results 4: The Clinical Implications

Aim 3: Identify how the biological findings may be integrated with clinical information in order to categorise patients of high and low-risk of progression.

6.1 Introduction

The molecular characterisation in the above chapters suggests a continuum of changes, punctuated by some key events, which correlate with advancing grade of disease. This gives us an increased understanding of the biology of progression, but the next step is to apply this information clinically to identify high-risk individuals. Whilst dysplasia is currently the gold-standard for identifying patients at high-risk of progressing to cancer, it is not perfect. Its diagnosis can be subjective, especially in the presence of inflammation. LGD is especially difficult to grade and there is significant inter-observer variability (Duits et al., 2015). Furthermore, although we treat all patients with confirmed dysplasia, only 11.6%/year of patients with LGD will progress to cancer (Phoa et al., 2014). So, the aim is to identify which combination of molecular features can best identify the high-risk individuals.

6.2 Decision tree

One way of categorising patients into risk groups would be a decision tree using multiple features. A decision tree has an advantage of being simple and easy to follow. We used a supervised learning algorithm (rpart package) with 16 features from genomic, expression and clinical data (Table 8) to grow a classification tree. The algorithm requires full data for each variable so smoking and body mass index could not be included, although they are known to be risk factors for BE. 87 samples had complete clinical information for inclusion. The genomic and expression features were all calculated using the same methods as used in the earlier chapters (see Methods).

Genomic statistics	**Driver gene mutation status**	**Clinical features**	**Other events**	**Expression**
Total mutation burden	*TP53*	Age	Chromothripsis	Chromosomal instability signature
Total copy number change	*CDKN2A*	Maximum BE length/cm	Whole genome doubling	
Total SV count	*ERBB2*			
Ploidy	*GATA6*			
	ARID1A			
	SMARCA4			
	MUC6			

Table 8 Features used for decision tree design

SV = structural variant, BE = Barrett's oesophagus

The algorithm splits the data recursively, until the end criterion of dysplasia status is reached. At each split, it determines which individual variable will give the largest possible reduction in heterogeneity of the dysplasia status. The variable which best split the data was the total structural variant (SV) count with a cut-off of 79 SVs, calculated by the algorithm, followed by *TP53* mutation status, then the proportion of the chromosome instability (CIN) signature (cut-off >0.38) to classify the remaining samples (Figure 62). Previously, the focus has always been on the correlation of SNVs and CNAs with dysplasia grade (Gu et al., 2010). SVs have not been considered, likely because non-WGS sequencing methods could not call them. This algorithm found SVs to discriminate better between the grades than TMB or CNA. This fits with our genomic analysis which showed a good discrimination between ND and dysplastic and a gradual increase in SV burden with progression.

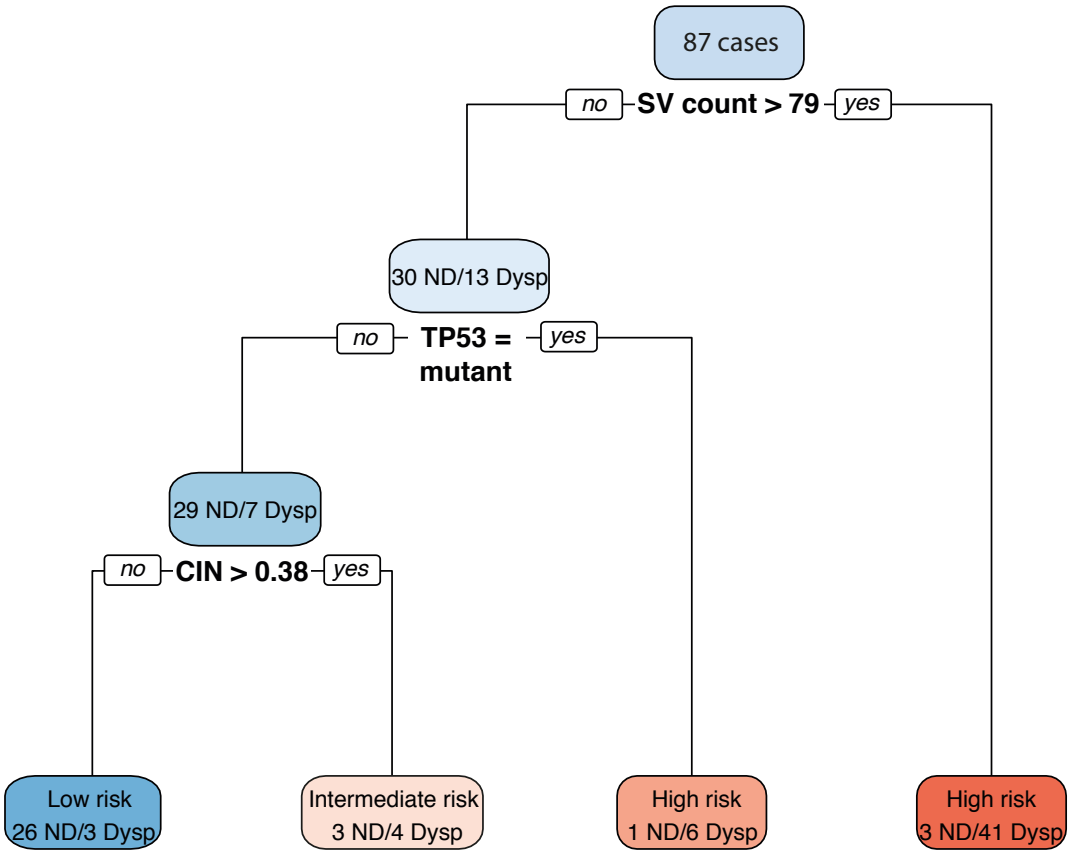


Figure 62 Decision tree model using rpart in R

The tool considered which of 16 features could best classify the samples by dysplasia status.SV = structural variant, CIN = chromosomal instability, ND = non-dysplastic, Dysp = dysplastic.

The CIN signature was able to further classify 4 dysplastic samples as being at risk. However, clinically, it would be advantageous to focus on one sequencing modality. We decided to reapply the clinical model using only SV count and *TP53* mutation status. This also meant that all 99 samples could be used, rather than only those with DNA and RNA sequenced. We then looked at all the other clinical and genomic features for the patients falling into each group to see if outliers displayed any distinguishing characteristics that were not in the initial model. One of the strengths of this cohort is that we have collated a comprehensive clinical annotation, including follow-up with future treatment and outcome.

Figure 63 summarises all of this information with a decision tree and heatmaps for each of the four classifications. Group 4 samples (low SV count, *TP53* wild type) would be considered low risk, groups 1-3 would be considered high-risk. The top bar of the heatmap denotes the overall grade of the patient (grey/yellow). As expected, the majority (87%) of the ND patients are in group 4. However, 8/42 in this group are dysplastic (3 LGD, 4 HGD, 1 IMC). We looked at these cases in more detail to see if there were other features to suggest that they were at risk.

Of the 8 outlier dysplastic cases in the low-risk group, notably, 4 were female, a far higher ratio than seen in BE. However, there was no statistically significant difference between the male:female ratios in group 1 (high SV count, *TP53* mutant) versus 4 (p value = 0.15, Fisher's exact test). None of the cases in the low risk group had whole genome duplication (WGD). The specific features of each of these outlier cases are detailed in Figure 64. All 8 patients were smokers but there were no other unifying features. Four of the 8 cases had a driver event in a gene other than CDKN2A (which is commonly mutated in NDBE). One case had a high mutation burden, greater than the median mutation burden in OAC (6.4 mutations/Mb) (Frankell et al., 2019). Potentially, these driver gene alterations could be the cause of progression or it may be another feature that we have not yet considered e.g. a gene affected by a specific SV. We noticed that 3 of the 8 dysplastic patients had an *ARID1A* alteration. Of the other 34 low risk patients, only 3 had an *ARID1A* alteration. Across the whole 99 pre-cancer BE patients there were 14 *ARID1A* alterations in total. 12 of which were in *TP53* wild type biopsies, suggesting a degree of mutual exclusivity. There was only a trend to significance: p value = 0.040 (Fisher's Exact Test) but this highlights the possibility of an alternative pathway to progression.

We reviewed the pathology consensuses of these cases. Firstly, of the frozen biopsies sequenced: seven had at least two of three pathology reviews agreeing on the presence of

dysplasia. However, one of the LGD cases, when further reviewed had a consensus of there not being any dysplasia within the frozen section. This may explain it falling into the low-risk category. The clinical FFPE block pathology reviews and follow-up were then checked for each of these cases. They all had at least 2 endoscopies on which dysplasia was confirmed and all had endoscopic treatment of the dysplasia, suggesting that these were definitely samples from patients who had progressed to dysplasia at the time they were taken.

Lastly, one possible explanation could have been that these samples were of lower cellularity and so the pipelines struggled to call the mutations. However, the computational cellularity was not significantly lower in these cases compared to other dysplastic cases (p value = 0.97, Wilcoxon Rank Sum). There was no significant difference between groups 1 and 4 for smoking status, age or length of BE.

One must consider the caveat of comparing to the gold standard of dysplasia. There is a possibility that the model has correctly identified these patients as low risk. The difficulty we have is that the patients were treated at the point of a diagnosis of dysplasia, so we do not know if they would have gone on to progress further.

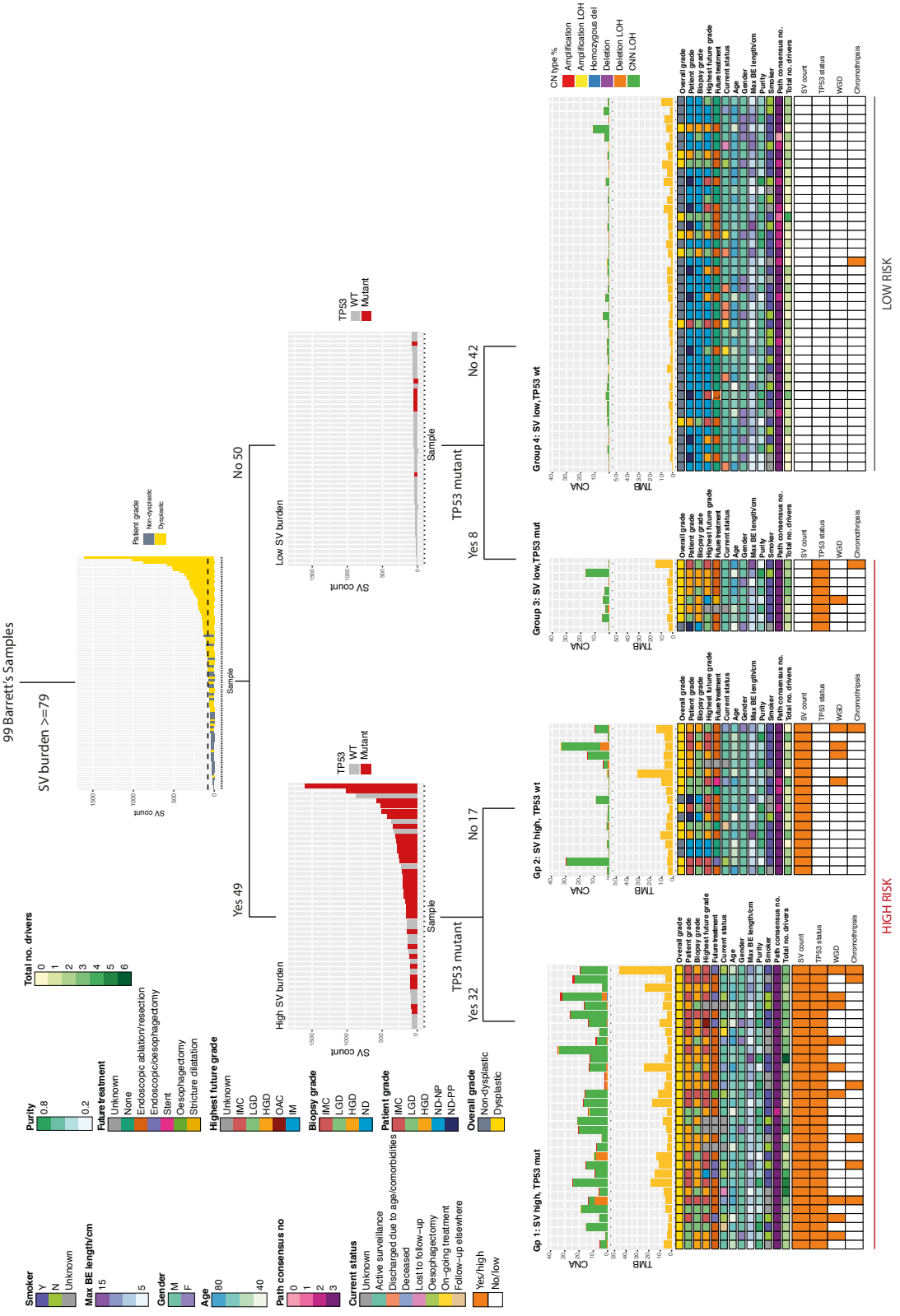


Figure 63 Classification decision tree using structural variant burden and *TP53* mutation status

Classification of 99 Barrett's oesophagus (BE) samples by structural variant (SV) burden followed by *TP53* mutation status (mut = mutant, wt = wildtype). SV count cut-off calculated by rpart algorithm. Cohort divides into 4 groups. Groups 1-3 considered high-risk and group 4 considered low risk. For each group, copy number aberrations (CNA) and total mutation burden (TMB) plotted. Clinical features and other genomic features displayed in heatmap annotation. WGD = whole genome duplication, IM = intestinal metaplasia, ND = Non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma, OAC = oesophageal adenocarcinoma.

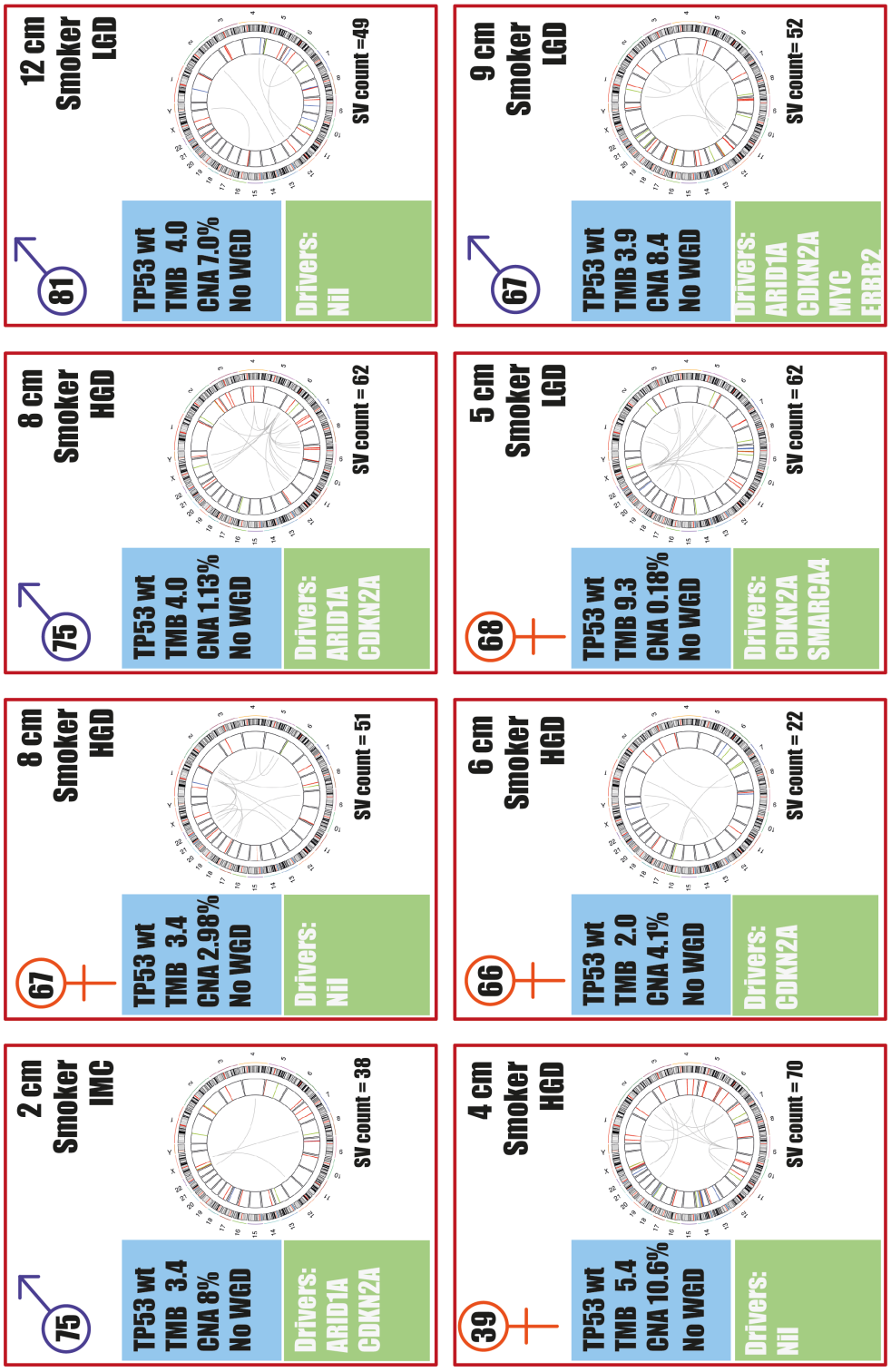


Figure 64 Clinical and genomic characteristics of dysplastic cases classified in the low risk group

Maximum length of BE segment given in cm. Circos plots show SVs for each case. TMB = total mutation burden, CNA = copy number aberrations, WGD = whole genome duplication, SV = structural variant, wt = wildtype, LGD = low grade dysplasia, HGD = high grade dysplasia, IMC = intramucosal carcinoma.

The five non-dysplastic outliers in high-risk groups 2 (high SV count, *TP53* wild type; 2 ND-NP, 2 ND-PP) and 3 (low SV count, *TP53* mutant; 1 ND-PP) were also investigated further (Figure 65). Of the two non-progressor cases, the 63-year-old male had 10 years of ND pathology on endoscopy. Four years before the sequenced biopsy and 6 after. This gives us confidence that he is truly a long-term non-progressor. He remains under surveillance. The 59-year-old male had a total of 13 years of surveillance. However, the biopsy which was sequenced was taken at his most recent surveillance in 2017 so there is the possibility that he may go on to progress in the future.

There were 3 pre-progressor cases categorised in the high-risk group. A 57-year-old male who progressed to LGD 2y 2m later. A 70-year-old male who actually had a clinical diagnosis of indefinite for dysplasia at the sequenced timepoint. He progressed to HGD on an endoscopy 1 year 3 months later. And a 55-year-old female, with a *TP53* mutation, who also had a diagnosis of indefinite for dysplasia on her endoscopy in 2008. She then had LGD 3.5 years later in 2012, followed by HGD 6 months after. Of course, with knowing the follow-up, we can see that these 3 cases were high-risk cases at this point, and perhaps, the decision tree correctly classified them. However, it classified the other ND-PP as low risk.

Overall, for a clinical test, overcalling a small proportion of cases, as above, is not such an issue, but the sensitivity is key so that patients with the disease are not missed. Using dysplasia grade as the gold-standard, SVs (>79 cut-off) and *TP53* mutation has a specificity of 87.2% and a sensitivity of 86.7% for its diagnosis. In comparison, when classifying the cohort using only *TP53* mutation, the specificity was excellent (98%), but the sensitivity dropped to 65%. This finding is in keeping with previous studies which used *TP53* mutation as a biomarker in BE (Ross-Innes et al., 2017). SVs alone had a similar specificity of 89.7%, and a sensitivity of 75%.

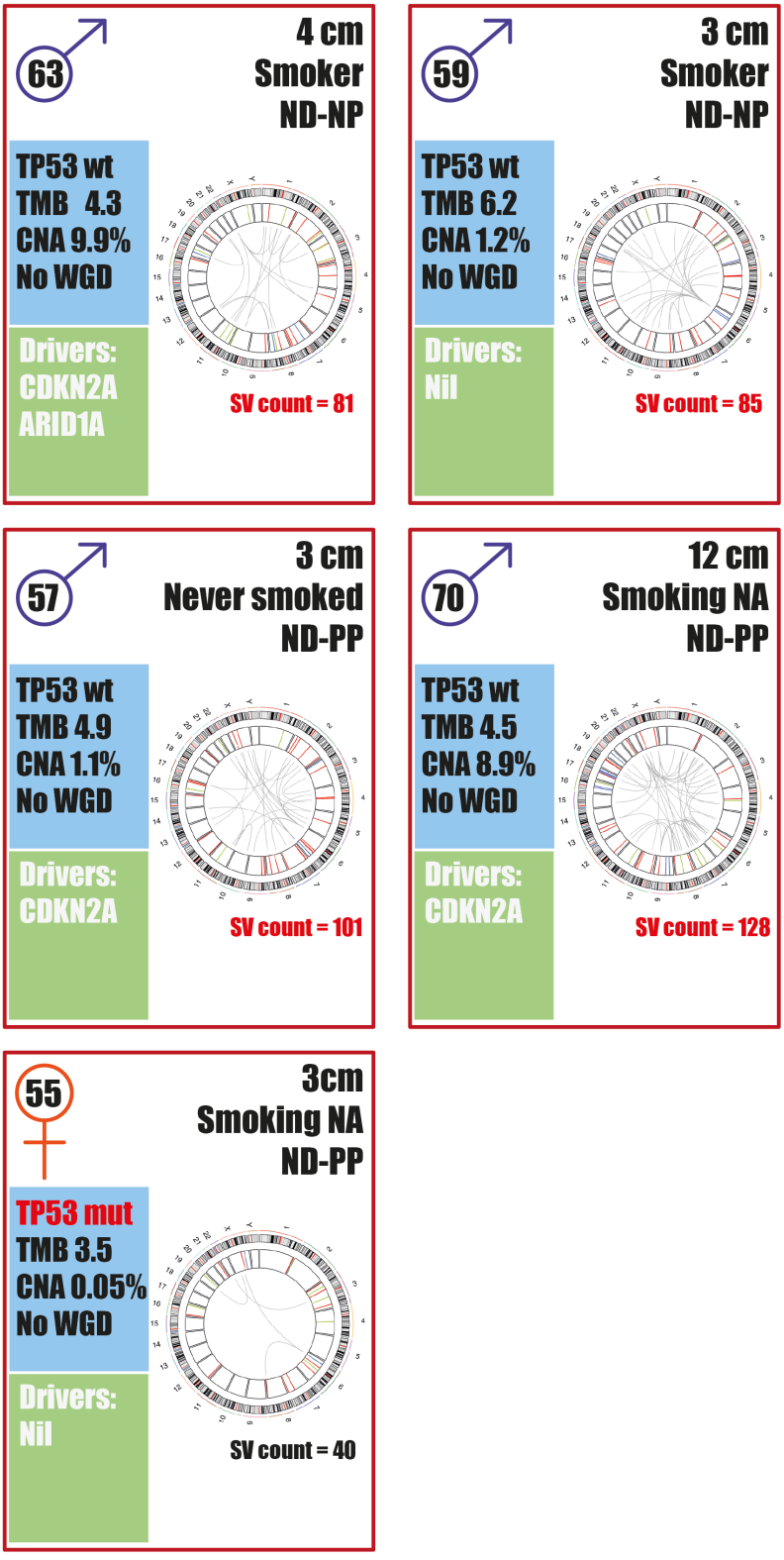


Figure 65 Clinical and genomic characteristics of non-dysplastic cases classified in the high-risk group

Maximum length of BE segment given in cm. Circos plots show SVs for each case. TMB = total mutation burden, CNA = copy number aberrations, WGD = whole genome duplication, SV = structural variant, wt = wildtype, NA = not available, ND = Non-dysplastic, NP = non-progressor, PP = pre-progressor.

As a comparison, we used the same criteria on the Trios BE (BE adjacent to cancer) and the adjacent cancers (Figure 66). The model only captured 24/47 of the BE cases (2 HGD, 4 LGD, 18 ND). The low risk group comprised of 1 HGD, 1 LGD, 1 indefinite for dysplasia and 20 ND. Whilst this highlights the findings in Results 1 and 2, that not all BE Trio samples act like their grade, the model only was able to classify 47% of the non-dysplastic samples. This raises the issue of the heterogeneity that we see in BE and a risk stratification test like this would not omit the need for multiple biopsies to be taken. In contrast, the model classified 46/47 (97.9%) of the OAC samples as high risk.

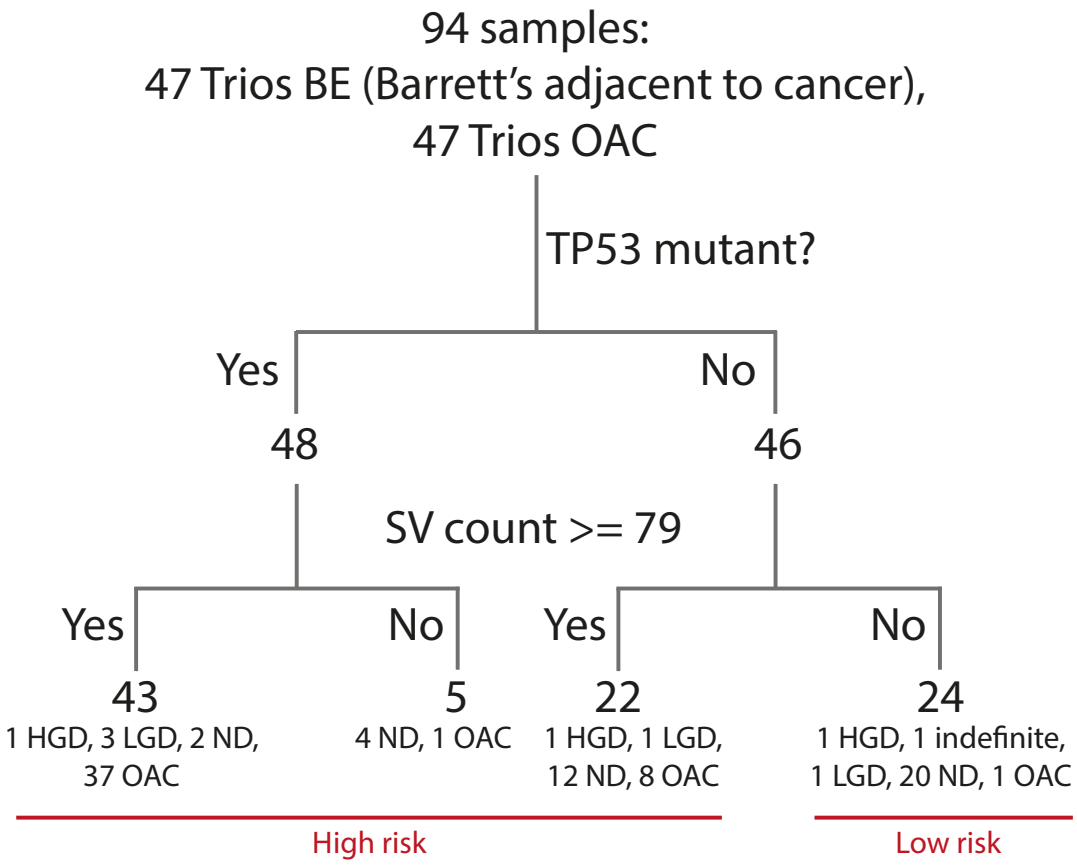


Figure 66 Classification of BE adjacent to cancer using the decision tree

Forty-seven BE samples adjacent to cancer (Trios) classified using the clinical decision model of TP53 mutation status and structural variant (SV) count. The pathology of each specific biopsy is given; however, all biopsies were taken at the cancer timepoint. ND = Non-dysplastic, NP = non-progressor, PP = pre-progressor, LGD = low grade dysplasia, HGD = high grade dysplasia.

6.3 Summary

We show that the structural variant count coupled with the *TP53* mutation status can be used risk stratify BE samples into high and low-risk groups. SV burden has not previously been evaluated.

A clinical risk stratification model for progression to HGD was recently published, assigning points for smoking (5 points), BE length (1 point/cm), male sex (9 points) and confirmed LGD on endoscopy (11 points) (Parasa et al., 2018). A score of >20 was considered to confer a high-risk of progression (2.1%/yr). All three of the LGD cases that fell into our low-risk class would score as high-risk in this model, mainly given the weighting of points for LGD. However, we assume that a LGD patient is misgrouped if they fall into the low risk category, even though they might never have progressed to HGD. This again highlights the difficulty with comparing to the current gold-standard of dysplasia. A caveat of their model is that it relies on both accurately sampling the LGD area on endoscopy and correctly diagnosing it. Both of which are difficult to do. It also cannot take into account the potential regression of LGD which was seen in 27.9% of patients in the SURF trial during the follow-up period (Phoa et al., 2014). This is an inherent confounder which is difficult to resolve now that patients are treated at the LGD stage and not followed-up for further progression.

From our 5 ND patients that fell into the higher risk groups, only 1 was high-risk using the Parasa et al. model but 3 were considered to be of intermediate risk. Adding smoking status to our model would have moved the dysplastic outliers into a higher risk category. However, 8 ND-NP and 5 ND-PP were also smokers in the low risk group so this would increase the false positive rate.

Whole genome sequencing for every patient with BE is not currently a viable option, mainly due to cost but also because it requires fresh/frozen tissue. However, if there are specific regions recurrently affected by SVs, it might be possible to then only sequence a portion of the genome and look for the breakpoints, especially if they are clustered. These regions are mainly fragile sites e.g. containing FHIT and WWOX and can be very large. FHIT is more than 1.5 million bp; WWOX 1.1 million. However, baits could be designed to pull out these whole regions for sequencing. Reducing the proportion of the genome being sequenced could facilitate a move to working with FFPE tissue: the current method of clinically preserving tissue.

Another problem is that the model still relies on the right tissue being sampled at endoscopy. We saw in our clonality analysis that there is a lot of heterogeneity between individual biopsies from a BE segment. The focus from here would be to test our model on a prospective cohort, more representative of the surveillance population, to see if the sensitivity and specificity hold on a less curated sample set.

If we are able to focus in on specific genomic regions, we could sequence to a higher depth. These methods could then be applied to much lower cellularity samples and, potentially, to non-endoscopic cell sampling devices such as the Cytosponge samples.

New sequencing methods, e.g. Nanopore technology (Oxford Nanopore Technologies, UK) may also make such an approach more practical. Nanopore sequencers can produce read lengths of up to 2Mb, without size selection, limited only by sample quality and library preparation method. In doing this it maps the breakpoints directly rather than using the unmapped reads, and provides a quick and easy method for SV detection (Sakamoto et al., 2019).

Discussion and future directions

We have performed an integrated genomic and transcriptomic analysis of 147 Barrett's oesophagus (BE) frozen samples across the grades. The aim was to gain an in depth understanding of the biology driving the progression of this disease so that this understanding could facilitate the future development of biomarkers for disease diagnosis and ultimately the earlier detection of oesophageal adenocarcinoma (OAC). A whole genome and transcriptome analysis of all the grades of BE has not previously been performed. Furthermore, whole genome sequencing (WGS) permits the analysis of structural variants which have never been considered at the pre-invasive stage. In addition, we compared the pre-cancer samples to BE sampled from adjacent to cancer in order to see whether BE adjacent to cancer is representative of pre-cancer BE or if the tumour development can have a local effect on the surrounding BE. We then went on to perform multilevel sequencing of 6 dysplastic BE cases to investigate the heterogeneity within segments. We used clonal and subclonal mutations and copy number alterations to understand degree of heterogeneity and the clonal evolution of BE. Finally, we considered how these findings could be applied clinically.

The strength of this cohort is the meticulous curation. The triple, independent consultant pathologist reviews gave us the highest possible certainty of the pathology of the biopsies that we sequenced. Patients had long-term follow-up and clear records of any treatment intervention. We were able to collate detailed, complete clinical information on each case. Good clinical annotation is rare in genomic studies. This allowed the integration of this information into progression models.

Creating such a well-defined cohort did present a number of challenges. We aimed to carefully select the frozen biopsies with high cellularity of the grades of interest. In cancer WGS studies, a pathological cellularity of 70% is commonly used. This percentage was not achievable for the BE because often the frozen research biopsy had not captured the dysplasia that the patient was known to have, resulting in the exclusion of most biopsies. We used a 30% cellularity cut-off and, importantly, have demonstrated that genomic alterations in BE can be confidently called at a depth of 50X WGS. We appreciate the disadvantage of only being able to cut and stain a single H&E from each frozen biopsy for diagnosis, due to the small size of the biopsies and the need for 2ug of DNA for WGS. It is possible that other

pathological grades were captured in the adjacent 2mm of tissue. However, for the main analysis, cases were analysed using the highest overall known grade of the patient, from FFPE, in order to negate the recognised difficulty in grading frozen tissue. We were unable to use microdissection in this study, again because of 2ug requirement for sequencing. Whilst this meant a lower cellularity, we avoided the possible introduction of artefact from DNA amplification. One point of which we must be aware is the possibility that we missed some real single nucleotide variants (SNVs) at low allele frequencies because of the relatively low depth of sequencing, the cellularity of the biopsies and the heterogeneity of BE with few clonal mutations. Early comparisons of Strelka and Mutect did show that Mutect called a slightly higher proportion of SNVs at very low allele frequencies (AF) (<0.1). But this must be weighed against the extra false positives that would have been called at these low AFs. We preferred to work with SNV calls that we could be confident of. An alternative would be to sequence to a higher depth of 100X but this would be very costly. For future work, a move towards using a new tissue preservation method, the PAXgene Tissue System (PreAnalytiX, Switzerland), could be used. This offers a similar quality of nucleic acid preservation as snap-freezing for whole genome and transcriptome sequencing, but maintains the tissue architecture for pathology review, as with FFPE tissue.

A high degree of heterogeneity between cases has previously been demonstrated in OAC (Frankell et al., 2019) and in small numbers of BE cases (Ross-Innes et al., 2015a). We observed in this large cohort that the heterogeneity of BE extended across the different genomic features but with a few key alterations occurring more frequently: particularly copy number alterations (CNAs) and mutations in driver genes. There were wide ranges in the numbers of SNVs, CNAs and structural variants (SVs) between samples, yet each feature appeared to be more of a gradual continuum when samples were ordered. We noticed that a specific genomic feature could dominate one sample, e.g. a dysplastic sample with a very high number of SVs but a low total mutation burden. In other cases, moderate numbers of each feature might be present. Driver mutations were then layered on top. Although de novo analyses did not discover any new drivers, this was not surprising given the size of the cohort. Using the known drivers in OAC, we were able to look back through the grades to see how early the alterations in these genes occurred. We noted a number of driver gene events to be specific to the dysplastic stage: *TP53* mutation, *GATA6* amplification, *ERBB2* amplification and *APC* mutation. All seen in >10% of dysplastic cases. *ERBB2* and *GATA6* demonstrated

a mutual exclusivity. *ARID1A* mutation, whilst seen in some of the non-dysplastic, non-progressor (ND-NP) cases, was mutually exclusive with *TP53* mutation.

Mutational signatures, in contrast, were less heterogeneous. They were preserved across the grades: seen from the early ND stage, both in non-progressors and pre-progressors. This suggests that they are set early with the formation of BE and the initial mutagenic environment to which the oesophagus is exposed. SV signatures, which have only very recently been described, were also preserved across the grades. There was a strong presence of signature 5 (dominated by translocations) throughout.

The possibility to analyse structural variants is one of the strengths of this analysis. So far only a basic analysis has been performed. The next step will be to look in more detail at the types of SV dominating each sample, rather than the signatures, and the genes affected by them. We would hope that further alterations in drivers might be identified. Furthermore, a specific look for the presence of mobile elements may help to define the numbers of translocations, as these can be confused by the SV caller Manta.

The transcriptomic analysis identified 475 differentially-expressed genes between ND and dysplastic cases. The genes downregulated from ND to dysplastic were also highly expressed in duodenum. Pathway analysis revealed the genes to have roles in intestinal development, fatty acid synthesis and lipid metabolism. A number of the genes were controlled upstream by *HNF4A*, a transcription factor important in the development of the intestines. These results support the theory that the intestinal metaplasia phenotype of the glandular epithelium is lost with progression. This has not previously been described on an expression level, despite being a phenotypic feature that has long been noted pathologically (Naini et al., 2016; Odze, 2006).

Pathway analyses of the genes upregulated in dysplasia highlighted the importance of the upregulation ERK/MAPK signalling pathway in progression. This pathway has previously been shown to be upregulated in BE cell lines with acid exposure (Jaiswal et al., 2006; Morgan et al., 2004; Souza et al., 2002) but this has not formerly been noted in large studies on human tissue.

Clinically, there is great interest in whether it is possible to identify patients with long-term indolent disease from those who are going to go on to progress, in order to guide the need for surveillance. In our pre-progressor (PP) cohort of ND samples, we did not see any differences in the genomic landscape compared to NP. Specifically, we did not see any *TP53*

mutations in this group, which has been claimed in other studies e.g. (Stachler et al., 2018). We appreciate that this was a small group in our cohort because it was very difficult to find frozen samples for cases who had long follow-up and progressed.

Previous WGS sequencing of BE has been performed on BE sampled from adjacent to cancer. But it has not been known if this tissue is representative of pre-cancer BE. We found that there were no significant differences in mutation burden or CNAs between the pre-cancer and adjacent to cancer BE, when ND and dysplastic were compared like for like. However, for several features, the ND samples appeared more similar to the dysplastic pre-cancer. There was a wider range of SV counts in the ND BE adjacent to cancer and two samples had evidence of chromothripsis. This was not seen in any of the ND pre-cancer cases. The spectrum of driver gene alterations in the ND BE adjacent to cancer were also more similar to the dysplastic samples, with alterations in *GATA6, TP53, SMAD4* and *KRAS*, albeit at low frequencies. However, this raises the issue that investigators need to be careful about extrapolating findings to the progression of BE when analysing samples taken at the cancer timepoint. It also supports preceding evidence of the local effect caused by the adjacent tumour.

Given the intralesional heterogeneity that has been observed in BE we wanted to look at some cases in more detail. We performed multilevel sequencing at 50X on 6 dysplastic cases. Analysis of the numbers of genomic alterations, signatures and driver gene alterations revealed differing degrees of heterogeneity between biopsies from a single case. Firstly, we found that there were some cases where all the biopsies from one case had similar numbers of SNVs, CNVs and SVs irrespective of the grade of each biopsy. For example, an IM biopsy in case 4 had a similar number of SNVs and SVs to its adjacent HGD sample 1cm distal to it. However, in other cases there was more variation between the biopsies e.g. with one biopsy with a much higher proportion of rearrangement than adjacent biopsies. In the analysis of signatures and drivers we saw further heterogeneity that did not always correlate with that seen in the numbers of SNVs, SVs or % of genome altered by CN. The clonality analysis, which brought the genomic features together to understand the relationships of clones to each other within the samples, further demonstrated the complexity of BE segments. For some cases we could construct phylogenetic trees with clear clonal ancestral clusters from which subclones developed. Yet, in others no clonal clusters were identified, with evidence of parallel evolution. We appreciate that these are 6 anecdotal cases and that more are required to make any defining statements about these phenomena.

It has been suggested that the diversity of a BE segment may be able to predict progression e.g. (Maley et al., 2006). We would like to go on to look at applying these measures of diversity to these samples. We have also defined a cohort of multilevel ND cases for a comparison in this clonality analysis which we will be sending for sequencing. Moving forward, it would be very interesting to be able to expand the multilevel cohort. Whole exome sequencing of small amounts FFPE cancer tissue has been optimised in our lab. If this could successfully be applied to the BE then the cohort could be expanded more cost-effectively. Using FFPE tissue would allow a much higher confidence of the grade that is being sequenced.

Overall, it seems that rather than a stepwise process with advancing pathological grades, molecularly there is more of a continuum. The current dysplasia status is categorical, in order to inform clinical management, but perhaps it is too simplistic. The phenotypic appearance is the end-point result of many molecular processes and the genomics suggest that there are many possible paths to progression. We hypothesise that molecular features accumulate over time until the combination of them is enough to tip the balance to progression. This is very important to understand when considering biomarkers for development: there is not one key defining genomic feature. A supervised learning algorithm identified SV count to be the variable which could best split the data into ND and dysplastic groups. It is interesting that SVs were better than CNAs, however technological advances are needed if SVs are to ever feature as biomarkers. It may be that further analysis will uncover a small number of recurrently rearranged regions which are enough to identify dysplasia. In this situation back-to-back baits could be designed to focus on sequencing the region. An alternative will be to work on a system to assign scores to cases for different genomic features based on their position within the continuum for each feature, e.g. SV count. The summation of these scores, and the definition of a threshold, could both identify cases with moderate levels of a number of features and cases with one overriding, dominant feature. Both which may be at a high risk of progression.

It is true that dysplasia is a reasonable biomarker when diagnosed by an expert GI pathologist but, clinically, GI pathologists are not available in every hospital. Potential alternatives to a molecular biomarker could be to centralise surveillance/pathology reviews or to work towards a machine-learning digital recognition of the grades. Machine-learning can now be used distinguish tumour from normal across tissue types on whole slide images (Campanella et al., 2019) and recently, correlation has been shown with genomic and transcriptomic

features (Fu et al., 2019, unpublished). However, so far work has mainly focussed on OAC and work on the pre-malignant stages is only just beginning to advance (Critchley-Thorne et al., 2017; El Hallani et al., 2015; Tomita et al., 2019).

Ultimately, we would like to move away from using phenotypic markers in the pathology for diagnosis and move to more quantitative molecular markers, with or without the addition of clinical features, to risk stratify patients. In our analysis, we have had to start by using dysplasia as a comparison and we consider LGD as progressed. But we know that not all patients with LGD would have progressed to cancer (11.6%/year) (Phoa et al., 2014). Now that we no longer get long-term follow-up on these LGD patients, because they are all treated, we cannot identify those that would have regressed/never progressed further. To move away from dysplasia entirely we ideally need to identify potential biomarkers and take the study back to FFPE material, where we can go further back in time with longer follow-up of LGD.

There are a number of future directions in which this project could be taken and endless potential analyses to continue on the cohort. Firstly, there are specific analyses that I would like to go on to perform but have been unable to date due to time limitations. Predominantly of the structural variants and how they compare in the multilevel cases. There are also more comparisons to do with the expression data in comparing different subgroups e.g. the non-progressors and pre-progressors, and also comparing NDBE to duodenum to see whether there are altered genes or pathways that distinguish it and could help to explain its potential for progression to cancer. Our lab has performed transcriptome sequencing on large numbers of tumours, and I would like to add a subset of these to the analysis to consider the changes in expression from the non-invasive IMC to the invasive cancer. A number of other future directions have also been mentioned in the discussions of the results chapters but, finally, the addition of methylation data would hugely strengthen this cohort and we are working towards this. This could then be integrated with the RNA using Multi-Omics Factor Analysis software (Argelaguet et al., 2018). For example, we may see changes in the methylation of genes in the ERK/MAPK to explain this upregulation.

This analysis was on a highly curated cohort and further work must look to see if specific alterations can be seen on any biopsies, irrespective of grade or cellularity. Again, the easiest way to do this would be to move to FFPE samples because of the wealth of banked diagnostic material accessible. It would also then be very interesting to track these genomic alterations back through time and ascertain when they first appeared.

Finally, the overarching aim of this project was to identify new biomarkers that could be applied to diagnose dysplasia and improve early detection of these lesions. The differential expression analysis has identified a number of candidates to take forward for testing, initially with IHC and then in panels, possibly combined with markers from other modalities.

Overall, we have shown that BE is highly heterogeneous across the molecular features of mutational patterns, copy-number aberrations (CNA) and structural variants (SVs). However, they reveal a gradual molecular continuum, rather than a stepwise progression through the pathological grades. This continuum is punctuated by certain key genomic alterations and we hypothesize that it is the accumulation of events that tips the balance to progression.

Bibliography

Agarwal, A., Polineni, R., Hussein, Z., Vigoda, I., Bhagat, T.D., Bhattacharyya, S., Maitra, A., and Verma, A. (2012). Role of epigenetic alterations in the pathogenesis of Barrett's esophagus and esophageal adenocarcinoma. Int. J. Clin. Exp. Pathol. *5*, 382–396.

Alexandrov SigProfiler, MATLAB Central File Exchange.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A. V, Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Ng, A.W., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E., Lopez-Bigas, N., et al. (2018). The Repertoire of Mutational Signatures in Human Cancer. BioRxiv 322859.

Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck, T.A., Simpson, J.T., Tonon, L., et al. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat. Commun. *6*, 10001.

Alvi, M.A., Liu, X., O'Donovan, M., Newton, R., Wernisch, L., Shannon, N.B., Shariff, K., di Pietro, M., Bergman, J.J., Ragunath, K., et al. (2013). DNA methylation as an adjunct to histopathology to detect prevalent, inconspicuous dysplasia and early-stage neoplasia in Barrett's esophagus. Clin Cancer Res *19*, 878–888.

Amin, M.B., Edge, S.B., and American Joint Committee on Cancer (2017). AJCC cancer staging manual (Springer).

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. *14*.

Bandla, S., Pennathur, A., Luketich, J.D., Beer, D.G., Lin, L., Bass, A.J., Godfrey, T.E., and Litle, V.R. (2012). Comparative genomics of esophageal adenocarcinoma and squamous cell carcinoma. Ann. Thorac. Surg. *93*, 1101–1106.

Bas Weusten, A., Bisschops, R., Coron, E., Dinis-Ribeiro, M., Dumonceau, J.-M., Esteban, J.-M., Hassan, C., Pech, O., Repici, A., Bergman, J., et al. (2017). Endoscopic management of Barrett's esophagus: European Society of Gastrointestinal Endoscopy (ESGE) Position

Statement.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. Proc. Natl. Acad. Sci. *104*, 20007–20012.

Bhat, S., Coleman, H.G., Yousef, F., Johnston, B.T., McManus, D.T., Gavin, A.T., and Murray, L.J. (2011). Risk of malignant progression in Barrett's esophagus patients: results from a large population-based study. J Natl Cancer Inst *103*, 1049–1057.

Bhat, S.K., McManus, D.T., Coleman, H.G., Johnston, B.T., Cardwell, C.R., McMenamin, Ú., Bannon, F., Hicks, B., Kennedy, G., Gavin, A.T., et al. (2015). Oesophageal adenocarcinoma and prior diagnosis of Barrett's oesophagus: a population-based study. Gut *64*, 20–25.

Bielski, C.M., Zehir, A., Penson, A. V., Donoghue, M.T.A., Chatila, W., Armenia, J., Chang, M.T., Schram, A.M., Jonsson, P., Bandlamudi, C., et al. (2018). Genome doubling shapes the evolution and prognosis of advanced cancers. Nat. Genet. *50*, 1189–1195.

Bird-Lieberman, E.L., Dunn, J.M., Coleman, H.G., Lao-Sirieix, P., Oukrif, D., Moore, C.E., Varghese, S., Johnston, B.T., Arthur, K., McManus, D.T., et al. (2012). Population-based study reveals new risk-stratification biomarker panel for Barrett's esophagus. Gastroenterology *143*, 927-35 e3.

Breton, J., Gage, M.C., Hay, A.W., Keen, J.N., Wild, C.P., Donnellan, C., Findlay, J.B.C., and Hardie, L.J. (2008). Proteomic Screening of a Cell Line Model of Esophageal Carcinogenesis Identifies Cathepsin D and Aldo-Keto Reductase 1C2 and 1B10 Dysregulation in Barrett's Esophagus and Esophageal Adenocarcinoma. J. Proteome Res. *7*, 1953–1962.

Burotto, M., Chiou, V.L., Lee, J.-M., and Kohn, E.C. (2014). The MAPK pathway across different malignancies: a new perspective. Cancer *120*, 3446–3456.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med.

Cancer Research UK Oesophageal cancer statistics.

Carter, S.L., Eklund, A.C., Kohane, I.S., Harris, L.N., and Szallasi, Z. (2006). A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat. Genet. *38*, 1043–1048.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.

Christensen, S., Roest, B. Van der, Besselink, N., Janssen, R., Boymans, S., Martens, J., Yaspo, M.-L., Priestley, P., Treatment, C. for P.C., Kuijk, E., et al. (2019). 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. BioRxiv 681262.

Cmero, M., Soon Ong, C., Yuan, K., Schröder, J., Mo, K., Evolution, P., Working Group, H., Corcoran, N.M., Papenfuss, T., Hovens, C.M., et al. SVclone: inferring structural variant cancer cell fraction Members of the PCAWG Evolution and Heterogeneity Working Group.

Coleman, H.G., Bhat, S.K., Murray, L.J., McManus, D.T., O'Neill, O.M., Gavin, A.T., Johnston, B.T., O'Neill, O.M., Gavin, A.T., and Johnston, B.T. (2014). Symptoms and endoscopic features at barrett's esophagus diagnosis: implications for neoplastic progression risk. Am J Gastroenterol *109*, 527–534.

Coleman, H.G., Xie, S.-H., and Lagergren, J. (2018). The Epidemiology of Esophageal Adenocarcinoma. Gastroenterology *154*, 390–405.

Corley, D.A., Kubo, A., Levin, T.R., Block, G., Habel, L., Rumore, G., Quesenberry, C., and Buffler, P. (2009). Race, ethnicity, sex and temporal differences in Barrett's oesophagus diagnosis: a large community-based study, 1994-2006. Gut *58*, 182–188.

Cortés-Ciriano, I., Lee, J.-K., Xi, R., Jain, D., Jung, Y.L., Yang, L., Gordenin, D., Klimczak, L.J., Zhang, C.-Z., Pellman, D.S., et al. (2018). Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. BioRxiv 333617.

Critchley-Thorne, R.J., Davison, J.M., Prichard, J.W., Reese, L.M., Zhang, Y., Repa, K., Li, J., Diehl, D.L., Jhala, N.C., Ginsberg, G.G., et al. (2017). A Tissue Systems Pathology Test Detects Abnormalities Associated with Prevalent High-Grade Dysplasia and Esophageal Cancer in Barrett's Esophagus. Cancer Epidemiol. Biomarkers Prev. *26*, 240–248.

Davelaar, A.L., Calpe, S., Lau, L., Timmer, M.R., Visser, M., Ten Kate, F.J., Parikh, K.B., Meijer, S.L., Bergman, J.J., Fockens, P., et al. (2015). Aberrant TP53 detected by combining immunohistochemistry and DNA-FISH improves Barrett's esophagus progression

prediction: a prospective follow-up study. Genes Chromosom. Cancer *54*, 82–90.

Davoli, T., Uno, H., Wooten, E.C., and Elledge, S.J. (2017). Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science (80-.). *355*, eaaf8399.

Dentro, S.C., Leshchiner, I., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A.G., Yu, K., Rubanova, Y., Macintyre, G., Vázquez-García, I., et al. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types Analysis of Whole Genomes Network.

Desai, T.K., Krishnan, K., Samala, N., Singh, J., Cluley, J., Perla, S., and Howden, C.W. (2012). The incidence of oesophageal adenocarcinoma in non-dysplastic Barrett's oesophagus: a meta-analysis. Gut *61*, 970–976.

Dilworth, M.P., Nieto, T., Stockton, J.D., Whalley, C.M., Tee, L., James, J.D., Noble, F., Underwood, T.J., Hallissey, M.T., Hejmadi, R., et al. (2019). Whole Genome Methylation Analysis of Nondysplastic Barrett Esophagus that Progresses to Invasive Cancer. Ann. Surg. *269*, 479–485.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dong, J., Buas, M.F., Gharahkhani, P., Kendall, B.J., Onstad, L., Zhao, S., Anderson, L.A., Wu, A.H., Ye, W., Bird, N.C., et al. (2018). Determining Risk of Barrett's Esophagus and Esophageal Adenocarcinoma Based on Epidemiologic Factors and Genetic Variants. Gastroenterology *154*, 1273-1281.e3.

Druliner, B.R., Wang, P., Bae, T., Baheti, S., Slettedahl, S., Mahoney, D., Vasmatzis, N., Xu, H., Kim, M., Bockol, M., et al. (2018). Molecular characterization of colorectal adenomas with and without malignancy reveals distinguishing genome, transcriptome and methylome alterations. Sci. Rep. *8*, 3161.

Duggan, S.P., Behan, F.M., Kirca, M., Zaheer, A., McGarrigle, S.A., Reynolds, J. V., Vaz, G.M.F., Senge, M.O., and Kelleher, D. (2016). The characterization of an intestine-like genomic signature maintained during Barrett's-associated adenocarcinogenesis reveals an NR5A2-mediated promotion of cancer cell survival. Sci. Rep. *6*, 32638.

Duits, L.C., Phoa, K.N., Curvers, W.L., Ten Kate, F.J., Meijer, G.A., Seldenrijk, C.A.,

Offerhaus, G.J., Visser, M., Meijer, S.L., Krishnadath, K.K., et al. (2015). Barrett's oesophagus patients with low-grade dysplasia can be accurately risk-stratified after histological review by an expert pathology panel. Gut *64*, 700–706.

Duits, L.C., Lao-Sirieix, P., Wolf, W.A., O'Donovan, M., Galeano-Dalmau, N., Meijer, S.L., Offerhaus, G.J.A., Redman, J., Crawte, J., Zeki, S., et al. (2019). A biomarker panel predicts progression of Barrett's esophagus to esophageal adenocarcinoma. Dis. Esophagus *32*.

Edelstein, Z.R., Bronner, M.P., Rosen, S.N., and Vaughan, T.L. (2009). Risk factors for Barrett's esophagus among patients with gastroesophageal reflux disease: a community clinic-based case-control study. Am. J. Gastroenterol. *104*, 834–842.

Elsner, M., Rauser, S., Maier, S., Schöne, C., Balluff, B., Meding, S., Jung, G., Nipp, M., Sarioglu, H., Maccarrone, G., et al. (2012). MALDI imaging mass spectrometry reveals COX7A2, TAGLN2 and S100-A10 as novel prognostic markers in Barrett's adenocarcinoma. J. Proteomics *75*, 4693–4704.

Eluri, S., Klaver, E., Duits, L.C., Jackson, S.A., Bergman, J.J., and Shaheen, N.J. (2018). Validation of a biomarker panel in Barrett's esophagus to predict progression to esophageal adenocarcinoma. Dis. Esophagus *31*.

Farshidfar, F., Zheng, S., Gingras, M.-C., Newton, Y., Shih, J., Robertson, A.G., Hinoue, T., Hoadley, K.A., Gibb, E.A., Roszik, J., et al. (2017). Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH -Mutant Molecular Profiles. Cell Rep. *18*, 2780–2794.

Fitzgerald, R. (2015). British Society of Gastroenterology Guidelines on the Diagnosis and Management of Barrett's Oesophagus - An Update.

Fitzgerald, R.C., Abdalla, S., Onwuegbusi, B.A., Sirieix, P., Saeed, I.T., Burnham, W.R., and Farthing, M.J.G. (2002). Inflammatory gradient in Barrett's oesophagus: implications for disease complications. Gut *51*, 316–322.

Fitzgerald, R.C., di Pietro, M., Ragunath, K., Ang, Y., Kang, J.Y., Watson, P., Trudgill, N., Patel, P., Kaye, P. V, Sanders, S., et al. (2014). British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus. Gut *63*, 7–42.

Frankel, A., Armour, N., Nancarrow, D., Krause, L., Hayward, N., Lampe, G., Smithers, B.M., and Barbour, A. (2014). Genome-wide analysis of esophageal adenocarcinoma yields specific copy number aberrations that correlate with prognosis. Genes, Chromosom. Cancer

53, 324–338.

Frankell, A.M., Jammula, S.G., Li, X., Contino, G., Killcoyne, S., Abbas, S., Perner, J., Bower, L., Devonshire, G., Ococks, E., et al. (2019). The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. Nat. Genet. *51*, 506–516.

Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., and Gerstung, M. (2019). Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. BioRChiv.

Galdiero, M.R., Bonavita, E., Barajon, I., Garlanda, C., Mantovani, A., and Jaillon, S. (2013). Tumor associated macrophages and neutrophils in cancer. Immunobiology *218*, 1402–1410.

Galipeau, P.C., Li, X., Blount, P.L., Maley, C.C., Sanchez, C.A., Odze, R.D., Ayub, K., Rabinovitch, P.S., Vaughan, T.L., and Reid, B.J. (2007). NSAIDs modulate CDKN2A, TP53, and DNA content risk for progression to esophageal adenocarcinoma. PLoS Med *4*, e67.

Gockel, I., Sgourakis, G., Lyros, O., Polotzek, U., Schimanski, C.C., Lang, H., Hoppo, T., and Jobe, B.A. (2011). Risk of lymph node metastasis in submucosal esophageal cancer: a review of surgically resected patients. Expert Rev. Gastroenterol. Hepatol. *5*, 371–384.

Gong, E.J., Kim, D.H., Ahn, J.Y., Jung, K.W., Lee, J.H., Choi, K.D., Song, H.J., Lee, G.H., Jung, H.-Y., Kim, H.S., et al. (2017). Comparison of long-term outcomes of endoscopic submucosal dissection and surgery for esophagogastric junction adenocarcinoma. Gastric Cancer *20*, 84–91.

Gu, J., Ajani, J.A., Hawk, E.T., Ye, Y., Lee, J.H., Bhutani, M.S., Hofstetter, W.L., Swisher, S.G., Wang, K.K., and Wu, X. (2010). Genome-wide catalogue of chromosomal aberrations in barrett's esophagus and esophageal adenocarcinoma: a high-density single nucleotide polymorphism array analysis. Cancer Prev Res *3*, 1176–1186.

Guan, B., Wang, T.-L., and Shih, I.-M. (2011). ARID1A, a Factor That Promotes Formation of SWI/SNF-Mediated Chromatin Remodeling, Is a Tumor Suppressor in Gynecologic Cancers. Cancer Res. *71*, 6718–6727.

El Hallani, S., Guillaud, M., Korbelik, J., and Marginean, E.C. (2015). Evaluation of Quantitative Digital Pathology in the Assessment of Barrett Esophagus–Associated

Dysplasia. Am. J. Clin. Pathol. *144*, 151–164.

Hamilton, J.P., Sato, F., Jin, Z., Greenwald, B.D., Ito, T., Mori, Y., Paun, B.C., Kan, T., Cheng, Y., Wang, S., et al. (2006). Reprimo Methylation Is a Potential Biomarker of Barrett's-Associated Esophageal Neoplastic Progression. Clin. Cancer Res. *12*, 6637–6642.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics *14*, 7.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Iorns, E., Martens-de Kemp, S.R., Lord, C.J., and Ashworth, A. (2009). CRK7 modifies the MAPK pathway and influences the response to endocrine therapy. Carcinogenesis *30*, 1696–1701.

Jaiswal, K., Lopez-Guzman, C., Souza, R.F., Spechler, S.J., and Sarosi, G.A. (2006). Bile salt exposure increases proliferation through p38 and ERK MAPK pathways in a non-neoplastic Barrett's cell line. Am. J. Physiol. Liver Physiol. *290*, G335–G342.

Jamal-Hanjani, M., Wilson, G.A., McGranahan, N., Birkbak, N.J., Watkins, T.B.K., Veeriah, S., Shafi, S., Johnson, D.H., Mitter, R., Rosenthal, R., et al. (2017). Tracking the evolution of non-small-cell lung cancer. N. Engl. J. Med. *376*, 2109–2121.

Janjigian, Y.Y., Bendell, J., Calvo, E., Kim, J.W., Ascierto, P.A., Sharma, P., Ott, P.A., Peltola, K., Jaeger, D., Evans, J., et al. (2018). CheckMate-032 Study: Efficacy and Safety of Nivolumab and Nivolumab Plus Ipilimumab in Patients With Metastatic Esophagogastric Cancer. J. Clin. Oncol. *36*, 2836–2844.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Kang, Y.-K., Boku, N., Satoh, T., Ryu, M.-H., Chao, Y., Kato, K., Chung, H.C., Chen, J.-S., Muro, K., Kang, W.K., et al. (2017). Nivolumab in patients with advanced gastric or gastro-oesophageal junction cancer refractory to, or intolerant of, at least two previous chemotherapy regimens (ONO-4538-12, ATTRACTION-2): a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet *390*, 2461–2471.

Karol Nowicki-Osuch, Lizhe Zhuang, Nils Eling, Sriganesh Jammula, Krishnaa T. Mahbubani, Annalise Katz-Summercorn, Anna Willbrey-Clark, Elo Madissoon, Massimiliano Di Pietro, Kerstin Meyer, Sarah Teichmann, Kourosh Saeb-Parsy, John C. Marioni, R.C.F. (2019). Cell profiling of the gastro-oesophageal junction informs pre-malignant and malignant transition.

Kastelein, F., Biermann, K., Steyerberg, E.W., Verheij, J., Kalisvaart, M., Looijenga, L.H., Stoop, H.A., Walter, L., Kuipers, E.J., Spaander, M.C., et al. (2013). Aberrant p53 protein expression is associated with an increased risk of neoplastic progression in patients with Barrett's oesophagus. Gut *62*, 1676–1683.

Katz-Summercorn, A., Anand, S., Ingledew, S., Huang, Y., Roberts, T., Galeano-Dalmau, N., O'Donovan, M., Liu, H., and Fitzgerald, R.C. (2017). Application of a multi-gene next-generation sequencing panel to a non-invasive oesophageal cell-sampling device to diagnose dysplastic Barrett's oesophagus. J. Pathol. Clin. Res. *3*, 258–267.

Kaye, P. V, Haider, S.A., Ilyas, M., James, P.D., Soomro, I., Faisal, W., Catton, J., Parsons, S.L., and Ragunath, K. (2009). Barrett's dysplasia and the Vienna classification: reproducibility, prediction of progression and impact of consensus reporting and p53 immunohistochemistry. Histopathology *54*, 699–712.

Kaz, A.M., Wong, C.-J., Luo, Y., Virgin, J.B., Washington, M.K., Willis, J.E., Leidner, R.S., Chak, A., and Grady, W.M. (2011). DNA methylation profiling in Barrett's esophagus and esophageal adenocarcinoma reveals unique methylation signatures and molecular subclasses. Epigenetics *6*, 1403–1412.

Krause, L., Nones, K., Loffler, K.A., Nancarrow, D., Oey, H., Tang, Y.H., Wayte, N.J., Patch, A.M., Patel, K., Brosda, S., et al. (2016). Identification of the CIMP-like subtype and aberrant methylation of members of the chromosomal segregation and spindle assembly pathways in esophageal adenocarcinoma. Carcinogenesis *37*, 356–365.

Krishnamoorthi, R., Singh, S., Ragunathan, K., Visrodia, K., Wang, K.K., Katzka, D.A., and Iyer, P.G. (2018). Factors Associated With Progression of Barrett's Esophagus: A Systematic Review and Meta-analysis. Clin. Gastroenterol. Hepatol. *16*, 1046-1055.e8.

Kuang, D.-M., Zhao, Q., Wu, Y., Peng, C., Wang, J., Xu, Z., Yin, X.-Y., and Zheng, L. (2011). Peritumoral neutrophils link inflammatory response to disease progression by fostering angiogenesis in hepatocellular carcinoma. J. Hepatol. *54*, 948–955.

Lavery, D.L., Nicholson, A.M., Poulsom, R., Jeffery, R., Hussain, A., Gay, L.J., Jankowski, J.A., Zeki, S.S., Barr, H., Harrison, R., et al. (2014). The stem cell organisation, and the proliferative and gene expression profile of Barrett's epithelium, replicates pyloric-type gastric glands. Gut *63*, 1854–1863.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Letouzé, E., Shinde, J., Renault, V., Couchy, G., Blanc, J.-F., Tubacher, E., Bayard, Q., Bacq, D., Meyer, V., Semhoun, J., et al. (2017). Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. Nat. Commun. *8*, 1315.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, X., Galipeau, P.C., Paulson, T.G., Sanchez, C.A., Arnaudo, J., Liu, K., Sather, C.L., Kostadinov, R.L., Odze, R.D., Kuhner, M.K., et al. (2014). Temporal and spatial evolution of somatic chromosomal alterations: a case-cohort study of Barrett's esophagus. Cancer Prev Res *7*, 114–127.

Lieberman, D.A., Oehlke, M., and Helfand, M. (1997). Risk factors for Barrett's esophagus in community-based practice. GORGE consortium. Gastroenterology Outcomes Research Group in Endoscopy. Am. J. Gastroenterol. *92*, 1293–1297.

Lin, E.W., Karakasheva, T.A., Hicks, P.D., Bass, A.J., and Rustgi, A.K. (2016). The tumor microenvironment in esophageal cancer. Oncogene *35*, 5337–5349.

Lind, A., Siersema, P.D., Kusters, J.G., Van der Linden, J.A.M., Knol, E.F., and Koenderman, L. (2012). The Immune Cell Composition in Barrett's Metaplastic Tissue Resembles That in Normal Duodenal Tissue. PLoS One *7*, e33899.

Van Loo, P., Nordgard, S.H., Lingjaerde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., et al. (2010). Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A *107*, 16910–16915.

Love, M.I., Huber, W., and Anders, S. (2014a). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Love, M.I., Huber, W., and Anders, S. (2014b). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Maag, J.L. V, Fisher, O.M., Levert-Mignon, A., Kaczorowski, D.C., Thomas, M.L., Hussey, D.J., Watson, D.I., Wettstein, A., Bobryshev, Y. V, Edwards, M., et al. (2017). Novel Aberrations Uncovered in Barrett's Esophagus and Esophageal Adenocarcinoma Using Whole Transcriptome Sequencing. Mol. Cancer Res. *15*, 1558–1569.

Maley, C.C., Galipeau, P.C., Li, X., Sanchez, C.A., Paulson, T.G., and Reid, B.J. (2004). Selectively Advantageous Mutations and Hitchhikers in Neoplasms. Cancer Res. *64*, 3414–3427.

Maley, C.C., Galipeau, P.C., Finley, J.C., Wongsurawat, V.J., Li, X., Sanchez, C.A., Paulson, T.G., Blount, P.L., Risques, R.-A., Rabinovitch, P.S., et al. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat. Genet. *38*, 468–473.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029-1041.e21.

Martinez, P., Timmer, M.R., Lau, C.T., Calpe, S., Sancho-Serra, M. del C., Straub, D., Baker, A.-M., Meijer, S.L., Kate, F.J.W. ten, Mallant-Hent, R.C., et al. (2016). Dynamic clonal equilibrium and predetermined cancer risk in Barrett's oesophagus. Nat. Commun. *7*, 12158.

Martinez, P., Mallo, D., Paulson, T.G., Li, X., Sanchez, C.A., Reid, B.J., Graham, T.A., Kuhner, M.K., and Maley, C.C. (2018). Evolution of Barrett's esophagus through space and time at single-crypt and whole-biopsy levels. Nat. Commun. *9*, 794.

McIntire, M.G., Soucy, G., Vaughan, T.L., Shahsafaei, A., and Odze, R.D. (2011). MUC2 Is a Highly Specific Marker of Goblet Cell Metaplasia in the Distal Esophagus and Gastroesophageal Junction. Am. J. Surg. Pathol. *35*, 1007–1013.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal

somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

Moons, L.M., Kusters, J.G., Bultman, E., Kuipers, E.J., van Dekken, H., Tra, W.M., Kleinjan, A., Kwekkeboom, J., van Vliet, A.H., and Siersema, P.D. (2005). Barrett's oesophagus is characterized by a predominantly humoral inflammatory response. J. Pathol. *207*, 269–276.

Moore, J.H., Lesser, E.J., Erdody, D.H., Natale, R.B., Orringer, M.B., and Beer, D.G. (1994). Intestinal differentiation and p53 gene alterations in Barrett's esophagus and esophageal adenocarcinoma. Int J Cancer *56*, 487–493.

Morgan, C., Alazawi, W., Sirieix, P., Freeman, T., Coleman, N., and Fitzgerald, R. (2004). In vitro acid exposure has a differential effect on apoptotic and proliferative pathways in a Barrett's adenocarcinoma cell line. Am. J. Gastroenterol. *99*, 218–224.

Murray, L., Sedo, A., Scott, M., McManus, D., Sloan, J.M., Hardie, L.J., Forman, D., and Wild, C.P. (2006). TP53 and progression from Barrett's metaplasia to oesophageal adenocarcinoma in a UK population cohort. Gut *55*, 1390–1397.

Murugaesu, N., Wilson, G.A., Birkbak, N.J., Watkins, T., McGranahan, N., Kumar, S., Abbassi-Ghadi, N., Salm, M., Mitter, R., Horswell, S., et al. (2015). Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. Cancer Discov. *5*, 821–831.

Naini, B. V., Souza, R.F., and Odze, R.D. (2016). Barrett's Esophagus: A Comprehensive and Contemporary Review for Pathologists. Am. J. Surg. Pathol. *40*, e45.

Nault, J.-C., Couchy, G., Balabaud, C., Morcrette, G., Caruso, S., Blanc, J.-F., Bacq, Y., Calderaro, J., Paradis, V., Ramos, J., et al. (2017). Molecular Classification of Hepatocellular Adenoma Associates With Risk Factors, Bleeding, and Malignant Transformation. Gastroenterology *152*, 880-894.e6.

Neshat, K., Sanchez, C.A., Galipeau, P.C., Blount, P.L., Levine, D.S., Joslyn, G., and Reid, B.J. (1994). p53 Mutations in Barrett's adenocarcinoma and high-grade dysplasia. Gastroenterology *106*, 1589–1595.

Newell, F., Patel, K., Gartside, M., Krause, L., Brosda, S., Aoude, L.G., Loffler, K.A., Bonazzi, V.F., Patch, A.-M., Kazakoff, S.H., et al. (2019). Complex structural rearrangements are present in high-grade dysplastic Barrett's oesophagus samples. BMC Med. Genomics *12*, 31.

NICE (2014). Endoscopic radiofrequency ablation for Barrett's oesophagus with low-grade dysplasia or no dysplasia.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al. (2012a). The Life History of 21 Breast Cancers. Cell *149*, 994–1007.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012b). Mutational Processes Molding the Genomes of 21 Breast Cancers. Cell *149*, 979–993.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole genome sequences. Nature *534*, 47.

Niu, B., Ye, K., Zhang, Q., Lu, C., Xie, M., McLellan, M.D., Wendl, M.C., and Ding, L. (2014). MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics *30*, 1015–1016.

Noble, F., Mellows, T., McCormick Matthews, L.H., Bateman, A.C., Harris, S., Underwood, T.J., Byrne, J.P., Bailey, I.S., Sharland, D.M., Kelly, J.J., et al. (2016). Tumour infiltrating lymphocytes correlate with improved survival in patients with oesophageal adenocarcinoma. Cancer Immunol. Immunother. *65*, 651–662.

Nones, K., Waddell, N., Wayte, N., Patch, A.M., Bailey, P., Newell, F., Holmes, O., Fink, J.L., Quinn, M.C., Tang, Y.H., et al. (2014). Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. Nat Commun *5*, 5224.

O'Neill, J.R., Pak, H.-S., Pairo-Castineira, E., Save, V., Paterson-Brown, S., Nenutil, R., Vojtěšek, B., Overton, I., Scherl, A., and Hupp, T.R. (2017). Quantitative Shotgun Proteomics Unveils Candidate Novel Esophageal Adenocarcinoma (EAC)-specific Proteins. Mol. Cell. Proteomics *16*, 1138.

Odze, R.D. (2006). Diagnosis and grading of dysplasia in Barrett's oesophagus. J. Clin. Pathol. *59*, 1029–1038.

Othman, M.O., Lee, J.H., and Wang, K. (2019). Clinical Practice Update on the Utility of Endoscopic Submucosal Dissection in T1b Esophageal Cancer: Expert Review. Clin. Gastroenterol. Hepatol. *17*, 2161–2166.

Owen, R.P., White, M.J., Severson, D.T., Braden, B., Bailey, A., Goldin, R., Wang, L.M.,

Ruiz-Puig, C., Maynard, N.D., Green, A., et al. (2018). Single cell RNA-seq reveals profound transcriptional similarity between Barrett's oesophagus and oesophageal submucosal glands. Nat. Commun. *9*, 4261.

Parasa, S., Vennalaganti, S., Gaddam, S., Vennalaganti, P., Young, P., Gupta, N., Thota, P., Cash, B., Mathur, S., Sampliner, R., et al. (2018). Development and Validation of a Model to Determine Risk of Progression of Barrett's Esophagus to Neoplasia. Gastroenterology *154*, 1282-1289.e2.

Paterson, A.L., Shannon, N.B., Lao-Sirieix, P., Ong, C.-A.J., Peters, C.J., O'Donovan, M., and Fitzgerald, R.C. (2013). A systematic approach to therapeutic target selection in oesophago-gastric cancer. Gut *62*, 1415–1424.

Paulson, T.G., Maley, C.C., Li, X., Li, H., Sanchez, C.A., Chao, D.L., Odze, R.D., Vaughan, T.L., Blount, P.L., and Reid, B.J. (2009). Chromosomal instability and copy number alterations in Barrett's esophagus and esophageal adenocarcinoma. Clin Cancer Res *15*, 3305–3314.

Peng, D., Sheta, E.A., Powell, S.M., Moskaluk, C.A., Washington, K., Goldknopf, I.L., and El-Rifai, W. (2008). Alterations in Barrett's-related adenocarcinomas: A proteomic approach. Int. J. Cancer *122*, 1303–1310.

Phoa, K.N., van Vilsteren, F.G.I., Weusten, B.L. a. M.B.L.A.M., Bisschops, R., Schoon, E.J., Ragunath, K., Fullarton, G., Di Pietro, M., Ravi, N., Visser, M., et al. (2014). Radiofrequency Ablation vs Endoscopic Surveillance for Patients With Barrett Esophagus and Low-Grade Dysplasia A Randomized Clinical Trial. Jama-Journal Am. Med. Assoc. *311*, 1209–1217.

Pich, O., Muiños, F., Lolkema, M.P., Steeghs, N., Gonzalez-Perez, A., and Lopez-Bigas, N. (2019). The mutational footprints of cancer therapies. BioRxiv 683268.

di Pietro, M., Boerwinkel, D.F., Shariff, M.K., Liu, X., Telakis, E., Lao-Sirieix, P., Walker, E., Couch, G., Mills, L., Nuckcheddy-Grant, T., et al. (2015). The combination of autofluorescence endoscopy and molecular biomarkers is a novel diagnostic tool for dysplasia in Barrett's oesophagus. Gut *64*, 49–56.

Del Portillo, A., Lagana, S.M., Yao, Y., Uehara, T., Jhala, N., Ganguly, T., Nagy, P., Gutierrez, J., Luna, A., Abrams, J., et al. (2015). Evaluation of Mutational Testing of Preneoplastic Barrett's Mucosa by Next-Generation Sequencing of Formalin-Fixed,

Paraffin-Embedded Endoscopic Samples for Detection of Concurrent Dysplasia and Adenocarcinoma in Barrett's Esophagus. J. Mol. Diagnostics *17*, 412–419.

Prevo, L.J., Sanchez, C.A., Galipeau, P.C., and Reid, B.J. (1999). p53-mutant clones and field effects in Barrett's esophagus. Cancer Res. *59*, 4784–4787.

Qu, X., Sandmann, T., Frierson, H., Fu, L., Fuentes, E., Walter, K., Okrah, K., Rumpel, C., Moskaluk, C., Lu, S., et al. (2016). Integrated genomic analysis of colorectal cancer progression reveals activation of EGFR through demethylation of the EREG promoter. Oncogene *35*, 6403–6415.

Que, J., Garman, K.S., Souza, R.F., and Spechler, S.J. (2019). Pathogenesis and Cells of Origin of Barrett's Esophagus. Gastroenterology *157*, 349-364.e1.

Reid, B.J., Haggitt, R.C., Rubin, C.E., and Rabinovitch, P.S. (1987). Barrett's esophagus: Correlation between flow cytometry and histology in detection of patients at risk for adenocarcinoma. Gastroenterology *93*, 1–11.

Reid, B.J., Blount, P.L., Rubin, C.E., Levine, D.S., Haggitt, R.C., and Rabinovitch, P.S. (1992). Flow-cytometric and histological progression to malignancy in Barrett's esophagus: Prospective endoscopic surveillance of a cohort. Gastroenterology *102*, 1212–1219.

Reid, B.J., Levine, D.S., Longton, G., Blount, P.L., and Rabinovitch, P.S. (2000). Predictors of progression to cancer in Barrett's esophagus: baseline histology and flow cytometry identify low- and high-risk patient subsets. Am. J. Gastroenterol. *95*, 1669–1676.

Reid, B.J., Prevo, L.J., Galipeau, P.C., Sanchez, C.A., Longton, G., Levine, D.S., Blount, P.L., and Rabinovitch, P.S. (2001). Predictors of progression in Barrett's esophagus II: baseline 17p (p53) loss of heterozygosity identifies a patient subset at increased risk for neoplastic progression. Am J Gastroenterol *96*, 2839–2848.

Reis, C.A., David, L., Correa, P., Carneiro, F., de Bolós, C., Garcia, E., Mandel, U., Clausen, H., and Sobrinho-Simões, M. (1999). Intestinal metaplasia of human stomach displays distinct patterns of mucin (MUC1, MUC2, MUC5AC, and MUC6) expression. Cancer Res. *59*, 1003–1007.

Riegman, P.H., Vissers, K.J., Alers, J.C., Geelen, E., Hop, W.C., Tilanus, H.W., and van Dekken, H. (2001). Genomic alterations in malignant transformation of Barrett's esophagus. Cancer Res *61*, 3164–3170.

Rogerson, C., Britton, E., Withey, S., Hanley, N., Ang, Y.S., and Sharrocks, A.D. (2019).

Identification of a primitive intestinal transcription factor network shared between esophageal adenocarcinoma and its precancerous precursor state. Genome Res. *29*, 723–736.

Ronkainen, J., Aro, P., Storskrubb, T., Johansson, S., Lind, T., Bolling–Sternevald, E., Vieth, M., Stolte, M., Talley, N.J., and Agréus, L. (2005). Prevalence of Barrett's Esophagus in the General Population: An Endoscopic Study. Gastroenterology *129*, 1825–1831.

Ross-Innes, C.S., Becq, J., Warren, A., Cheetham, R.K., Northen, H., O'Donovan, M., Malhotra, S., di Pietro, M., Ivakhno, S., He, M., et al. (2015a). Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. Nat. Genet. *47*, 1038–1046.

Ross-Innes, C.S., Debiram-Beecham, I., O'Donovan, M., Walker, E., Varghese, S., Lao-Sirieix, P., Lovat, L., Griffin, M., Ragunath, K., Haidry, R., et al. (2015b). Evaluation of a minimally invasive cell sampling device coupled with assessment of trefoil factor 3 expression for diagnosing Barrett's esophagus: a multi-center case-control study. PLoS Med *12*, e1001780.

Ross-Innes, C.S., Chettouh, H., Achilleos, A., Galeano-Dalmau, N., Debiram-Beecham, I., MacRae, S., Fessas, P., Walker, E., Varghese, S., Evan, T., et al. (2017). Risk stratification of Barrett's oesophagus using a non-endoscopic sampling method coupled with a biomarker panel: a cohort study. Lancet Gastroenterol. Hepatol. *2*, 23–31.

Sakamoto, Y., Xu, L., Seki, M., Yokoyama, T.T., Kasahara, M., Kashima, Y., Ohashi, A., Shimada, Y., Motoi, N., Tsuchihara, K., et al. (2019). Long read sequencing reveals a novel class of structural aberrations in cancers: identification and characterization of cancerous local amplifications. BioRxiv 620047.

Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811–1817.

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: the Pathway Interaction Database. Nucleic Acids Res. *37*, D674–D679.

Schneider, P.M., Casson, A.G., Levin, B., Garewal, H.S., Hoelscher, A.H., Becker, K., Dittler, H.J., Cleary, K.R., Troster, M., Siewert, J.R., et al. (1996). Mutations of p53 in Barrett's esophagus and Barrett's cancer: a prospective study of ninety-eight cases. J Thorac Cardiovasc Surg *111*, 323.

Secrier, M., Li, X., De Silva, N., Eldridge, M.D., Contino, G., Bornschein, J., Macrae, S., Grehan, N., O'Donovan, M., Miremadi, A., et al. (2016). Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. Nat. Genet. *48*, 1131–1141.

Shaheen, N.J., Sharma, P., Overholt, B.F., Wolfsen, H.C., Sampliner, R.E., Wang, K.K., Galanko, J.A., Bronner, M.P., Goldblum, J.R., Bennett, A.E., et al. (2009). Radiofrequency Ablation in Barrett's Esophagus with Dysplasia. N. Engl. J. Med. *360*, 2277–2288.

Shaheen, N.J., Falk, G.W., Iyer, P.G., and Gerson, L.B. (2016). ACG Clinical Guideline: Diagnosis and Management of Barrett's Esophagus. Am. J. Gastroenterol. *111*, 30–50.

Sikkema, M., Kerkhof, M., Steyerberg, E.W., Kusters, J.G., van Strien, P.M., Looman, C.W., van Dekken, H., Siersema, P.D., and Kuipers, E.J. (2009). Aneuploidy and overexpression of Ki67 and p53 as markers for neoplastic progression in Barrett's esophagus: a case-control study. Am J Gastroenterol *104*, 2673–2680.

Skacel, M., Petras, R.E., Rybicki, L.A., Gramlich, T.L., Richter, J.E., Falk, G.W., and Goldblum, J.R. (2002). p53 expression in low grade dysplasia in Barrett's esophagus: correlation with interobserver agreement and disease progression. Am J Gastroenterol *97*, 2508–2513.

Somja, J., Demoulin, S., Roncarati, P., Herfs, M., Bletard, N., Delvenne, P., and Hubert, P. (2013). Dendritic Cells in Barrett's Esophagus Carcinogenesis. Am. J. Pathol. *182*, 2168–2179.

Sommerer, F., Vieth, M., Markwarth, A., Röhrich, K., Vomschloß, S., May, A., Ell, C., Stolte, M., Hengge, U.R., Wittekind, C., et al. (2004). Mutations of BRAF and KRAS2 in the development of Barrett's adenocarcinoma. Oncogene *23*, 554–558.

Souza, R.F., Shewmake, K., Terada, L.S., and Spechler, S.J. (2002). Acid exposure activates the mitogen-activated protein kinase pathways in Barrett's esophagus. Gastroenterology *122*, 299–307.

Stachler, M.D., Taylor-Weiner, A., Peng, S., McKenna, A., Agoston, A.T., Odze, R.D., Davison, J.M., Nason, K.S., Loda, M., Leshchiner, I., et al. (2015). Paired exome analysis of Barrett's esophagus and adenocarcinoma. Nat. Genet. *47*, 1047–1055.

Stachler, M.D., Camarda, N.D., Deitrick, C., Kim, A., Agoston, A.T., Odze, R.D., Hornick, J.L., Nag, A., Thorner, A.R., Ducar, M., et al. (2018). Detection of Mutations in Barrett's

Esophagus Before Progression to High-Grade Dysplasia or Adenocarcinoma. Gastroenterology *155*, 156–167.

Streitz, J.M., Madden, M.T., Marimanikkuppam, S.S., Krick, T.P., Salo, W.L., and Aufderheide, A.C. (2005). Analysis of protein expression patterns in Barrett's esophagus using Maldi mass spectrometry, in search of malignancy biomarkers. Dis. Esophagus *18*, 170–176.

Sugimura, K., Miyata, H., Tanaka, K., Takahashi, T., Kurokawa, Y., Yamasaki, M., Nakajima, K., Takiguchi, S., Mori, M., and Doki, Y. (2015). High infiltration of tumor-associated macrophages is associated with a poor response to chemotherapy and poor prognosis of patients undergoing neoadjuvant chemotherapy for esophageal cancer. J. Surg. Oncol. *111*, 752–759.

Tamborero, D., Rubio-Perez, C., Muiños, F., Sabarinathan, R., Piulats, J.M., Muntasell, A., Dienstmann, R., Lopez-Bigas, N., and Gonzalez-Perez, A. (2018). A Pan-cancer Landscape of Interactions between Solid Tumors and Infiltrating Immune Cell Populations. Clin. Cancer Res. *24*, 3717–3728.

TCGA (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519–525.

TCGA, Kim, J.J., Bowlby, R., Mungall, A.J., Robertson, A.G., Odze, R.D., Cherniack, A.D., Shih, J., Pedamallu, C.S., Cibulskis, C., et al. (2017). Integrated genomic characterization of oesophageal carcinoma. Nature *541*, 169–175.

Tecchio, C., Scapini, P., Pizzolo, G., and Cassatella, M.A. (2013). On the cytokines produced by human neutrophils in tumors. Semin. Cancer Biol. *23*, 159–170.

Teixeira, V.H., Pipinikas, C.P., Pennycuick, A., Lee-Six, H., Chandrasekharan, D., Beane, J., Morris, T.J., Karpathakis, A., Feber, A., Breeze, C.E., et al. (2019). Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. Nat. Med. *25*, 517–525.

Thrumurthy, S.G., Chaudry, M.A., Thrumurthy, S.S.D., and Mughal, M. (2019). Oesophageal cancer: risks, prevention, and diagnosis. BMJ *366*, l4373.

Togashi, Y., Shitara, K., and Nishikawa, H. (2019). Regulatory T cells in cancer immunosuppression — implications for anticancer therapy. Nat. Rev. Clin. Oncol. *16*, 356–371.

Tomita, N., Abdollahi, B., Wei, J., Ren, B., Suriawinata, A., and Hassanpour, S. (2019). Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides. JAMA Netw. Open *2*, e1914645.

Tomkova, M., Tomek, J., Kriaucionis, S., and Schuster-Böckler, B. (2018). Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. *19*, 129.

Tran Janco, J.M., Lamichhane, P., Karyampudi, L., and Knutson, K.L. (2015). Tumor-infiltrating dendritic cells in cancer pathogenesis. J. Immunol. *194*, 2985–2991.

Ünal, E.B., Uhlitz, F., and Blüthgen, N. (2017). A compendium of ERK targets. FEBS Lett. *591*, 2607–2615.

Varagunam, M., Brand, C., Cromwell, D., Maynard, N., Crosby, T., Michalowski, J., and Napper, R. (2017). National Oesophago-Gastric Cancer Audit 2017.

Varghese, S., Newton, R., Ross-Innes, C.S., Lao-Sirieix, P., Krishnadath, K.K., O'Donovan, M., Novelli, M., Wernisch, L., Bergman, J., and Fitzgerald, R.C. (2015). Analysis of dysplasia in patients with Barrett's esophagus based on expression pattern of 90 genes. Gastroenterology *149*, 1511-1518.e5.

Vennalaganti, P., Kanakadandi, V., Goldblum, J.R., Mathur, S.C., Patil, D.T., Offerhaus, G.J., Meijer, S.L., Vieth, M., Odze, R.D., Shreyas, S., et al. (2017). Discordance Among Pathologists in the United States and Europe in Diagnosis of Low-Grade Dysplasia for Patients With Barrett's Esophagus. Gastroenterology *152*, 564-570.e4.

Wani, S., Puli, S.R., Shaheen, N.J., Westhoff, B., Slehria, S., Bansal, A., Rastogi, A., Sayana, H., and Sharma, P. (2009). Esophageal adenocarcinoma in Barrett's esophagus after endoscopic ablative therapy: A meta-analysis and systematic review. Am. J. Gastroenterol. *104*, 502–513.

Weaver, J.M.J., Ross-Innes, C.S., Shannon, N., Lynch, A.G., Forshew, T., Barbera, M., Murtaza, M., Ong, C.-A.A.J., Lao-Sirieix, P., Dunning, M.J., et al. (2014). Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. Nat Genet *46*, 837–843.

Westhoff, B., Brotze, S., Weston, A., McElhinney, C., Cherian, R., Mayo, M.S., Smith, H.J., and Sharma, P. (2005). The frequency of Barrett's esophagus in high-risk patients with chronic GERD. Gastrointest. Endosc. *61*, 226–231.

Weston, A.P., Banerjee, S.K., Sharma, P., Tran, T.M., Richards, R., and Cherian, R. (2001). p53 protein overexpression in low grade dysplasia (LGD) in Barrett's esophagus: immunohistochemical marker predictive of progression. Am J Gastroenterol *96*, 1355–1362.

Wong, D.J., Barrett, M.T., Stoger, R., Emond, M.J., and Reid, B.J. (1997). p16INK4a promoter is hypermethylated at a high frequency in esophageal adenocarcinomas. Cancer Res *57*, 2619–2622.

Xu, E., Gu, J., Hawk, E.T., Wang, K.K., Lai, M., Huang, M., Ajani, J., and Wu, X. (2013). Genome-wide methylation analysis shows similar patterns in Barrett's esophagus and esophageal adenocarcinoma. Carcinogenesis *34*, 2750–2756.

Younes, M., Ertan, A., Lechago, L. V, Somoano, J.R., and Lechago, J. (1997). p53 Protein accumulation is a specific marker of malignant potential in Barrett's metaplasia. Dig Dis Sci *42*, 697–701.

Zagari, R.M., Fuccio, L., Wallander, M.-A., Johansson, S., Fiocca, R., Casanova, S., Farahmand, B.Y., Winchester, C.C., Roda, E., and Bazzoli, F. (2008). Gastro-oesophageal reflux symptoms, oesophagitis and Barrett's oesophagus in the general population: the Loiano-Monghidoro study. Gut *57*, 1354–1359.

ZHANG, W., and LIU, H.T. (2002). MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Res. *12*, 9–18.

Zhao, J., Chang, A.C., Li, C., Shedden, K.A., Thomas, D.G., Misek, D.E., Manoharan, A.P., Giordano, T.J., Beer, D.G., and Lubman, D.M. (2007). Comparative Proteomics Analysis of Barrett Metaplasia and Esophageal Adenocarcinoma Using Two-dimensional Liquid Mass Mapping. Mol. Cell. Proteomics *6*, 987–999.

Zhou, Z., Kalatskaya, I., Russell, D., Marcon, N., Cirocco, M., Krzyzanowski, P.M., Streutker, C., Liang, H., Litle, V.R., Godfrey, T.E., et al. (2019a). Combined EsophaCap cytology and MUC2 immunohistochemistry for screening of intestinal metaplasia, dysplasia and carcinoma. Clin. Exp. Gastroenterol. *12*, 219–229.

Zhou, Z., Kalatskaya, I., Russell, D., Marcon, N., Cirocco, M., Krzyzanowski, P., Streutker, C., Liang, H., Litle, V., Godfrey, T., et al. (2019b). Combined EsophaCap cytology and MUC2 immunohistochemistry for screening of intestinal metaplasia, dysplasia and carcinoma. Clin. Exp. Gastroenterol. *Volume 12*, 219–229.

Supplementary

Supplementary Table 1 Significantly up and down-regulated genes in dysplastic versus non-dysplastic BE

Only gene with at least a 3-fold change (log2 fold change >1.58 or <-1.58) and adjusted p values < 0.05 are included.

Gene ID	Base Mean Counts	Log2 Fold Change	p value	p adj
KRTAP9-7	5.95	30.00	1.92E-22	3.34E-18
SOHLH1	4.72	7.85	1.36E-15	0.000186927
CST1	18.08	6.97	1.36E-15	1.13E-06
RP11-169F17.1	34.72	6.95	1.36E-15	7.41E-07
LGSN	19.25	6.03	1.36E-15	4.73E-07
CFHR4	2.70	4.59	1.36E-15	0.017143393
DSPP	5.35	4.46	1.36E-15	0.000401334
IGFL4	17.11	4.39	1.36E-15	1.10E-09
POTEE	41.14	4.36	1.36E-15	3.65E-08
POTEF	314.58	4.33	1.36E-15	6.17E-11
PYDC2	4.40	4.28	1.36E-15	6.98E-06
LHX1	5.50	4.15	1.36E-15	6.69E-05
DMP1	4.81	4.12	1.36E-15	0.030825147
SERPINA3	12.54	3.91	1.36E-15	3.01E-07
DKK2	5.69	3.90	1.36E-15	8.47E-08
CGA	6.32	3.76	1.36E-15	0.000292122
IGF2BP3	107.12	3.74	3.62E-20	2.10E-16
FETUB	5.67	3.68	1.36E-15	0.043087653
CXCL5	217.01	3.67	1.36E-15	5.59E-06
FGF20	9.91	3.66	1.36E-15	4.77E-05
CYP4F31P	13.75	3.64	1.36E-15	6.02E-06
NT5DC4	3.25	3.59	1.36E-15	5.96E-07
OBP2B	15.43	3.58	1.36E-15	4.82E-06
MMP3	86.48	3.52	1.36E-15	6.67E-11
AQP9	11.29	3.48	1.36E-15	9.59E-09
TH	3.39	3.47	1.36E-15	7.66E-05
NOTUM	313.19	3.45	1.36E-15	1.08E-08
OBP2A	4.90	3.45	1.36E-15	0.0006538
ADAMTS20	2.17	3.42	1.36E-15	0.004718436
MMP1	350.83	3.37	1.07E-19	4.66E-16
CTCFL	6.11	3.35	1.36E-15	0.003310892
FGF19	2.48	3.33	1.36E-15	0.002464023
STAC2	4.74	3.32	1.36E-15	1.74E-05
LAIR2	10.35	3.31	1.36E-15	2.64E-08

COL22A1	54.25	3.19	1.36E-15	1.65E-07
BIRC7	4.86	3.18	1.36E-15	2.80E-06
KCNK9	6.53	3.13	1.36E-15	1.34E-07
BRDT	1.44	3.13	1.36E-15	0.01138016
IBSP	1.52	3.12	1.36E-15	0.02483184
WT1	1.29	3.07	1.36E-15	0.034606888
CXCR1	13.39	3.07	1.36E-15	4.11E-06
WNT11	26.24	3.06	1.36E-15	1.24E-07
TREM1	15.83	3.06	1.36E-15	1.85E-07
ALPK2	15.92	3.04	1.36E-15	3.11E-09
CRABP1	2.43	3.03	1.36E-15	0.002136444
IL8	272.21	2.99	9.03E-17	1.36E-13
DUSP27	170.30	2.99	1.36E-15	9.78E-06
CACNG8	18.11	2.98	1.36E-15	1.08E-08
RDH8	3.31	2.97	1.36E-15	0.002165161
CALML5	56.47	2.92	1.36E-15	0.007416955
TBX4	14.39	2.91	1.36E-15	3.05E-06
POTEI	114.41	2.90	1.36E-15	2.32E-08
LY6G6C	9.47	2.90	1.36E-15	7.96E-06
PCDH15	6.35	2.88	1.36E-15	0.001387434
POTEJ	66.17	2.86	1.36E-15	1.19E-07
ZNF716	2.17	2.82	1.36E-15	0.025122442
CSF3	14.69	2.82	1.36E-15	1.68E-07
SLCO1A2	4.44	2.82	1.36E-15	2.09E-05
HS3ST5	9.50	2.81	1.36E-15	4.37E-08
GPR3	6.58	2.81	1.36E-15	1.12E-10
TCF24	7.60	2.79	1.36E-15	3.57E-08
EGR3	403.55	2.74	1.36E-15	6.30E-10
NPFFR1	24.76	2.70	1.36E-15	8.99E-10
KIRREL2	1.92	2.70	1.36E-15	9.59E-05
PAEP	2.87	2.68	1.36E-15	0.030921005
SCN2A	77.42	2.68	1.36E-15	2.88E-05
TNS4	923.54	2.67	1.36E-15	1.37E-09
LPA	10.78	2.65	1.36E-15	0.004761608
SULT4A1	23.12	2.63	1.36E-15	0.015900028
ASTL	7.00	2.58	1.36E-15	2.15E-06
PPBP	29.84	2.57	1.36E-15	8.38E-08
CCDC166	1.30	2.56	1.36E-15	0.008238724
OSM	27.53	2.53	1.36E-15	3.00E-07
C3orf72	2.95	2.53	1.36E-15	0.001083247
SLCO1B1	4.80	2.50	1.36E-15	0.025584121
NXPH1	4.85	2.49	1.36E-15	0.000230687

MOBP	2.06	2.48	1.36E-15	0.000519851
MMP13	22.58	2.48	1.36E-15	2.03E-05
DAZL	1.17	2.46	1.36E-15	0.007324411
ULBP1	6.22	2.46	1.36E-15	1.23E-06
DUSP2	185.35	2.45	1.36E-15	3.15E-11
IL22	1.45	2.44	1.36E-15	0.013402714
RPRM	5.99	2.43	1.36E-15	0.002994114
C11orf85	1.45	2.43	1.36E-15	0.001012144
GALNT13	4.10	2.42	1.36E-15	0.032038949
IZUMO2	1.03	2.41	1.36E-15	0.019174137
NKD1	273.11	2.41	1.36E-15	4.63E-09
TNNI2	25.81	2.40	1.36E-15	3.87E-07
IGFBP1	2.53	2.39	1.36E-15	0.005558038
MUC16	149.09	2.39	1.36E-15	0.000507691
TGM6	5.44	2.39	1.36E-15	0.02822288
POU6F2	20.50	2.37	1.36E-15	0.002789051
GLT1D1	9.35	2.35	1.36E-15	3.37E-06
SLCO1B7	2.41	2.35	1.36E-15	0.016114473
VNN3	15.87	2.35	1.36E-15	1.92E-07
AP000695.1	1.68	2.35	1.36E-15	0.000526003
DRD1	54.65	2.34	1.36E-15	1.73E-07
FCGR3B	30.26	2.33	1.36E-15	4.04E-06
FAM155A	79.68	2.33	1.36E-15	1.49E-09
GOLGA6B	3.74	2.31	1.36E-15	2.51E-05
CXCR5	2.83	2.31	1.36E-15	0.004394933
GOLGA6D	3.64	2.30	1.36E-15	4.20E-05
CPZ	2.17	2.29	1.36E-15	0.00170178
GOLGA6C	3.89	2.29	1.36E-15	2.75E-05
FPR1	17.77	2.29	1.36E-15	1.40E-06
NOS1	7.50	2.28	1.36E-15	0.005985348
FAM178B	7.06	2.28	1.36E-15	0.001102428
MYH4	0.93	2.28	1.36E-15	0.04332023
OPTC	1.79	2.27	1.36E-15	0.019224146
GSDMA	31.90	2.27	1.36E-15	1.35E-05
PTX4	5.34	2.26	1.36E-15	0.002273945
KRT83	1.55	2.26	1.36E-15	0.033691004
GNGT1	12.15	2.25	1.36E-15	0.000960885
ONECUT1	11.52	2.25	1.36E-15	0.003210619
CHI3L1	37.47	2.25	1.36E-15	3.34E-05
FAM71F1	4.45	2.25	1.36E-15	0.020876637
OR13J1	1.16	2.24	1.36E-15	0.011594728
IL24	10.04	2.24	1.36E-15	8.04E-05

RP11-553A10.1	11.02	2.23	1.36E-15	3.32E-05
EGR4	24.74	2.23	1.36E-15	2.49E-07
SYT1	47.71	2.22	1.36E-15	2.47E-07
PATE1	1.41	2.22	1.36E-15	0.021576265
CSF3R	102.17	2.21	1.36E-15	6.69E-09
KIAA1199	1607.95	2.20	1.36E-15	3.99E-10
FBN3	9.41	2.20	1.36E-15	0.013260447
FOXL2	1.58	2.20	1.36E-15	0.039078631
ALB	12.65	2.18	1.36E-15	0.001912217
AC091801.1	4.85	2.18	1.36E-15	0.000311505
AC079354.1	20.31	2.18	1.36E-15	1.10E-05
POTEH	2.27	2.18	1.36E-15	0.006359277
PATE3	2.05	2.17	1.36E-15	0.010750761
ARMC3	27.11	2.17	1.36E-15	1.68E-08
LIPN	4.19	2.17	1.36E-15	0.00978692
CTNNA3	52.87	2.16	1.36E-15	4.57E-05
DLX4	5.27	2.16	1.36E-15	3.51E-07
SLC12A3	2.15	2.16	1.36E-15	0.003930119
GOLGA6A	5.19	2.15	1.36E-15	3.14E-06
FPR2	5.84	2.15	1.36E-15	0.000682928
PGC	8494.09	2.14	1.36E-15	0.000691813
SPDYC	14.32	2.14	1.36E-15	3.19E-06
C4orf50	2.16	2.14	1.36E-15	0.000543894
NPHS1	5.52	2.14	1.36E-15	6.15E-05
C3	1153.49	2.14	1.36E-15	1.46E-05
CILP2	10.03	2.13	1.36E-15	4.04E-06
LILRA2	3.99	2.13	1.36E-15	0.000101106
SNAP91	2.25	2.13	1.36E-15	0.018520105
COL20A1	2.96	2.11	1.36E-15	0.000492341
CACNA1S	0.89	2.10	1.36E-15	0.016425409
MUC19	4.20	2.10	1.36E-15	0.016359523
CD300E	8.41	2.10	1.36E-15	0.000376851
CXXC11	0.94	2.10	1.36E-15	0.016208458
TREML2	6.66	2.09	1.36E-15	0.000105428
OLR1	11.65	2.09	1.36E-15	3.17E-05
OR2B6	3.78	2.09	1.36E-15	0.000777512
PKD2L1	2.84	2.09	1.36E-15	3.16E-05
NPY5R	1.28	2.09	1.36E-15	0.011957235
ESM1	10.25	2.09	1.36E-15	2.71E-05
C20orf202	8.22	2.08	1.36E-15	5.08E-06
DRD2	4.85	2.08	1.36E-15	5.68E-05

MAT1A	2.77	2.08	1.36E-15	0.007584259
HORMAD1	59.57	2.08	1.36E-15	6.61E-05
GALNT16	40.46	2.07	1.36E-15	8.16E-07
FOXC2	2.91	2.07	1.36E-15	0.005773195
CCDC19	21.08	2.07	1.36E-15	8.65E-10
CRYAA	0.81	2.06	1.36E-15	0.026635693
SERPINE1	133.36	2.05	1.36E-15	3.51E-07
FNDC7	1.15	2.05	1.36E-15	0.004846854
INHBA	71.56	2.05	1.36E-15	8.38E-07
LGALS12	1.58	2.05	1.36E-15	0.005807401
IL17C	4.78	2.04	1.36E-15	8.67E-05
PCDHA4	26.49	2.04	1.36E-15	0.000713756
RP11-204N11.1	1.56	2.04	1.36E-15	0.003026554
IL11	17.95	2.04	1.36E-15	9.15E-06
BARX1	13.43	2.04	1.36E-15	0.004666353
FXYD7	1.42	2.03	1.36E-15	0.001586432
CXCL1	608.03	2.03	5.27E-18	1.31E-14
CAMKV	2.57	2.02	1.36E-15	0.000955993
TRIM72	4.14	2.02	1.36E-15	0.000334983
TNNT2	55.41	2.02	1.36E-15	1.68E-07
WNT7B	46.09	2.01	1.36E-15	2.42E-05
RASGEF1A	58.23	2.01	1.36E-15	4.68E-05
HSD3B2	0.74	2.00	1.36E-15	0.042784445
SCGB3A2	0.93	2.00	1.36E-15	0.035908529
POU4F1	1.65	2.00	1.36E-15	0.037086319
GPR132	109.32	2.00	1.36E-15	9.10E-10
CXCL2	641.78	2.00	1.36E-15	1.12E-10
PROK1	2.99	1.99	1.36E-15	0.005045279
IL6	19.50	1.99	1.36E-15	0.002909253
STRA6	47.31	1.98	1.36E-15	2.92E-06
RP1	18.09	1.97	1.36E-15	0.000272465
CLDN6	1.41	1.97	1.36E-15	0.016862073
LPO	2.50	1.97	1.36E-15	0.024132142
EMR3	6.64	1.97	1.36E-15	0.000879918
KCNK15	28.54	1.96	1.36E-15	2.85E-08
AADACL2	22.27	1.96	1.36E-15	8.38E-07
USP17L22	2.42	1.96	1.36E-15	0.004008162
TPPP3	614.33	1.95	1.36E-15	2.80E-09
PATE2	1.99	1.95	1.36E-15	0.035329671
FAM71A	1.86	1.94	1.36E-15	0.027153159
AL603965.1	24.21	1.94	1.36E-15	1.79E-05

FBXO2	75.92	1.94	1.36E-15	3.19E-06
MMP10	12.46	1.94	1.36E-15	0.000205046
SALL4	41.47	1.93	1.36E-15	3.96E-08
USP17L15	4.45	1.92	1.36E-15	0.001767798
RIMS4	13.68	1.91	1.36E-15	9.83E-06
TBC1D29	2.23	1.91	1.36E-15	0.001846042
TDRD1	22.22	1.90	1.36E-15	6.75E-05
SELL	47.49	1.90	1.36E-15	1.39E-05
AL161645.2	1.89	1.90	1.36E-15	0.025260101
LAMP5	15.87	1.90	1.36E-15	0.000602364
USP17L11	0.94	1.90	1.36E-15	0.033095008
C1orf180	2.18	1.89	1.36E-15	0.021154843
NEUROD2	20.53	1.89	1.36E-15	0.000213872
PTGS2	270.77	1.89	1.36E-15	2.56E-07
ACTL6B	2.14	1.89	1.36E-15	0.003392331
UBE2U	5.56	1.89	1.36E-15	0.037500474
SYCP2	328.68	1.89	1.36E-15	1.96E-05
TNR	3.83	1.89	1.36E-15	0.006163775
PATE4	5.87	1.88	1.36E-15	0.000357854
CSTL1	5.28	1.88	1.36E-15	0.000264608
THBS2	1241.31	1.88	1.36E-15	3.96E-07
LRP8	212.73	1.88	1.36E-15	7.10E-09
INSL4	1.97	1.87	1.36E-15	0.044150281
TNNI3	2.59	1.87	1.36E-15	0.00089081
C6ORF165	0.82	1.86	1.36E-15	0.024625011
TMEM78	1.55	1.86	1.36E-15	0.00848786
TSNAX-DISC1	1.19	1.86	1.36E-15	0.025311987
NFE2	8.24	1.86	1.36E-15	0.000215409
DOC2A	26.21	1.86	1.36E-15	7.58E-06
RUFY4	27.93	1.86	1.36E-15	7.34E-06
MSLNL	48.88	1.86	1.36E-15	1.04E-05
CHRDL2	16.72	1.86	1.36E-15	1.54E-06
LRP2	2.23	1.85	1.36E-15	0.016206189
PLA2G2E	1.30	1.85	1.36E-15	0.044968679
CCK	66.14	1.85	1.36E-15	0.004255003
DEFA1B	5.72	1.85	1.36E-15	0.026466273
UCN2	5.48	1.85	1.36E-15	0.000605261
FGF23	1.19	1.84	1.36E-15	0.023827092
FCAR	5.44	1.84	1.36E-15	0.000295657
FAM183B	0.93	1.84	1.36E-15	0.047817591
DEFA1	7.69	1.83	1.36E-15	0.020517583
PRED60	1.02	1.83	1.36E-15	0.021103513

NCR2	1.21	1.83	1.36E-15	0.015083827
PRSS50	2.33	1.82	1.36E-15	0.007268693
MMP9	34.98	1.82	1.36E-15	0.000156204
MROH2A	43.46	1.81	1.36E-15	2.77E-05
KIF26B	606.26	1.80	1.36E-15	2.00E-12
IGFN1	11.77	1.80	1.36E-15	0.000185214
PRSS55	10.48	1.79	1.36E-15	0.001153017
NRCAM	73.15	1.79	1.36E-15	3.27E-06
DLX2	1.07	1.79	1.36E-15	0.035448876
HAS1	4.35	1.79	1.36E-15	0.005032817
MAEL	21.52	1.78	1.36E-15	0.01800884
MEFV	26.95	1.78	1.36E-15	4.98E-06
CALCR	1.91	1.78	1.36E-15	0.034017947
FCN3	3.79	1.77	1.36E-15	0.00033386
MYO3B	23.64	1.77	1.36E-15	0.000172501
CR1	102.70	1.77	1.36E-15	0.00052628
USP17L20	2.83	1.77	1.36E-15	0.020872737
IFI6	2279.38	1.76	1.36E-15	2.25E-06
ADAM32	78.12	1.76	1.36E-15	4.30E-10
CCDC73	7.82	1.76	1.36E-15	0.000665192
FKBP10	885.92	1.75	1.36E-15	8.02E-07
IL19	6.45	1.75	1.36E-15	0.031833436
MFI2	289.42	1.75	1.36E-15	2.52E-08
PF4V1	1.68	1.75	1.36E-15	0.030204259
HAP1	28.97	1.75	1.36E-15	0.000127574
CTD-2215E18.1	2.38	1.75	1.36E-15	0.005877006
PEG10	79.87	1.75	1.36E-15	0.000105475
TMEM88B	1.41	1.74	1.36E-15	0.024478661
CCL19	26.30	1.74	1.36E-15	0.001310365
IFI44L	460.81	1.74	1.36E-15	2.49E-06
NLRP12	9.04	1.74	1.36E-15	1.12E-05
CTRB2	5.20	1.73	1.36E-15	0.000490886
LILRA1	6.08	1.73	1.36E-15	0.000765311
BTBD19	301.43	1.73	1.36E-15	6.74E-09
LEMD1	26.00	1.73	1.36E-15	5.30E-05
EYA4	21.71	1.73	1.36E-15	0.00020117
GEM	369.03	1.73	1.36E-15	2.74E-11
TEX101	2.05	1.73	1.36E-15	0.020579901
C17orf96	60.57	1.72	1.36E-15	1.10E-09
KLK14	7.45	1.72	1.36E-15	0.000876727
VWA3A	18.50	1.72	1.36E-15	3.48E-06

XIRP1	2.25	1.72	1.36E-15	0.018022323
CHST4	49.06	1.72	1.36E-15	2.14E-06
AC138655.1	1.15	1.72	1.36E-15	0.030109557
GNAT1	1.78	1.71	1.36E-15	0.025674319
FGF18	14.31	1.71	1.36E-15	1.19E-06
FNDC8	1.72	1.71	1.36E-15	0.0078328
ACTBL2	4.16	1.71	1.36E-15	0.007379203
TULP2	11.11	1.70	1.36E-15	2.07E-05
C6orf100	1.72	1.70	1.36E-15	0.007756737
SLC13A3	71.38	1.70	1.36E-15	1.53E-08
CYP2E1	107.32	1.70	1.36E-15	0.000228899
USP17L23	3.88	1.69	1.36E-15	0.010082908
C5AR1	75.96	1.69	1.36E-15	2.85E-09
TTC24	3.18	1.69	1.36E-15	0.001352187
NR4A3	373.53	1.69	1.36E-15	0.000201354
UPK3A	1.61	1.69	1.36E-15	0.012532855
OR2A14	2.08	1.69	1.36E-15	0.005199424
ANKRD45	11.91	1.68	1.36E-15	0.000591536
CLEC17A	13.77	1.68	1.36E-15	0.022872934
HEMGN	1.17	1.68	1.36E-15	0.015247062
IL31	1.50	1.68	1.36E-15	0.024759859
ABCA12	173.11	1.67	1.36E-15	0.001166208
ASB4	6.91	1.67	1.36E-15	0.023873143
RP11-758M4.1	4.80	1.67	1.36E-15	0.0051739
GAD1	57.06	1.67	1.36E-15	1.34E-07
PTP4A3	160.10	1.67	1.36E-15	1.07E-10
SIGLEC14	5.15	1.67	1.36E-15	0.002346837
POPDC3	4.02	1.66	1.36E-15	0.047172796
C4orf47	9.48	1.66	1.36E-15	7.03E-06
AC129492.6	1.27	1.66	1.36E-15	0.027570675
C1orf170	23.02	1.66	1.36E-15	1.84E-07
AKR1CL1	2.21	1.66	1.36E-15	0.022423945
ALOXE3	5.19	1.66	1.36E-15	0.027846449
ZDHHC11	686.89	1.66	1.36E-15	4.35E-08
LCN2	3496.68	1.66	1.36E-15	2.64E-05
SH2D5	6.12	1.65	1.36E-15	0.000113799
GAL	7.82	1.65	1.36E-15	0.018490922
NPW	13.92	1.65	1.36E-15	1.75E-06
ACAN	26.72	1.65	1.36E-15	0.001351188
RSPO4	3.45	1.65	1.36E-15	0.009276199
ANKRD1	1.28	1.65	1.36E-15	0.028910159

ATF3	2411.37	1.65	1.36E-15	1.07E-05
RASGRF1	9.33	1.65	1.36E-15	0.000349562
HPN	44.49	1.65	1.36E-15	0.000418321
ADAMTS4	167.04	1.65	1.36E-15	3.34E-05
NTSR1	85.42	1.64	1.36E-15	0.000128427
NTRK2	83.91	1.64	1.36E-15	0.000852219
SLC5A12	42.64	1.64	1.36E-15	0.000905185
CDKN2A	274.90	1.63	1.36E-15	0.004066277
SNAI1	23.48	1.63	1.36E-15	8.58E-08
C19orf26	18.02	1.63	1.36E-15	3.17E-06
PENK	7.75	1.62	1.36E-15	0.047125941
FRMPD1	10.06	1.62	1.36E-15	0.000155581
SLC6A3	4.06	1.62	1.36E-15	0.002057525
AGT	65.47	1.62	1.36E-15	0.000117612
AC104057.1	12.62	1.62	1.36E-15	0.001817801
XKR7	9.98	1.62	1.36E-15	0.042686079
FNDC1	16.66	1.61	1.36E-15	0.000990015
PAX5	55.32	1.61	1.36E-15	0.000770967
CALHM3	5.81	1.61	1.36E-15	0.003412775
ANGPT4	1.99	1.61	1.36E-15	0.008238724
GALR3	0.92	1.61	1.36E-15	0.033095008
RGS6	118.77	1.61	1.36E-15	0.000132262
CLEC4E	3.08	1.60	1.36E-15	0.024170784
GPR123	3.28	1.60	1.36E-15	0.008904558
LYZL4	1.23	1.59	1.36E-15	0.022403947
LILRA3	2.59	1.58	1.36E-15	0.017892613
CHRND	3.81	1.58	1.36E-15	0.00016344
ADAMTS14	239.31	1.58	1.36E-15	4.61E-07
CIDEC	344.43	-1.59	1.36E-15	6.05E-05
FMO6P	26.64	-1.59	1.36E-15	1.73E-05
PAK7	2.04	-1.59	1.36E-15	0.023248434
BTNL3	1150.78	-1.59	1.36E-15	6.77E-07
LIPF	8941.38	-1.60	1.36E-15	0.033768294
SLC16A9	717.86	-1.61	1.36E-15	5.08E-06
CHODL	9.97	-1.61	1.36E-15	0.000255641
TFF1	38740.32	-1.62	1.36E-15	3.95E-05
MME	235.18	-1.62	1.36E-15	0.004382842
SFTPC	1.60	-1.63	1.36E-15	0.046200897
SCNN1B	129.22	-1.63	1.36E-15	0.012183168
CYP4F2	437.82	-1.65	1.36E-15	3.56E-05
MOGAT2	660.38	-1.65	1.36E-15	2.69E-07
CNTFR	36.18	-1.65	1.36E-15	0.00146606

ANPEP	28494.27	-1.66	1.36E-15	0.000155581
LRRC19	421.59	-1.68	1.36E-15	1.45E-06
FAM189A1	31.31	-1.68	1.36E-15	0.007682982
PHGR1	6781.67	-1.68	1.36E-15	2.82E-07
KIAA1644	146.10	-1.68	1.36E-15	3.48E-06
TBX10	47.44	-1.68	1.36E-15	2.37E-05
ADRA1B	1.29	-1.70	1.36E-15	0.003965885
RGS2	3189.17	-1.70	1.36E-15	1.50E-06
HYAL4	8.88	-1.70	1.36E-15	0.00082885
MUC2	16721.91	-1.71	1.36E-15	8.19E-05
NEFM	7.66	-1.73	1.36E-15	2.40E-05
SOX11	1.32	-1.74	1.36E-15	0.036249185
AC138517.1	20.34	-1.74	1.36E-15	6.77E-09
ABCB11	12.84	-1.74	1.36E-15	1.03E-06
MLIP	27.91	-1.74	1.36E-15	3.18E-06
CDA	78.77	-1.74	1.36E-15	2.27E-05
APOBEC1	168.87	-1.75	1.36E-15	1.81E-07
C19orf69	9.26	-1.76	1.36E-15	0.013167283
AQP10	29.01	-1.76	1.36E-15	0.003764773
EDN3	225.11	-1.76	1.36E-15	2.06E-06
RGR	1.22	-1.77	1.36E-15	0.014994553
CASR	23.04	-1.77	1.36E-15	0.013039787
NDST4	11.69	-1.77	1.36E-15	2.81E-05
AGXT	33.13	-1.77	1.36E-15	1.50E-06
NAALADL1	81.59	-1.77	1.36E-15	3.02E-07
SERPINA6	8.59	-1.78	1.36E-15	0.000755262
TMEM179	0.90	-1.79	1.36E-15	0.008028811
CDX1	1448.46	-1.79	1.36E-15	5.43E-08
TMEM229A	124.90	-1.80	1.36E-15	1.23E-05
KLK15	9.08	-1.80	1.36E-15	5.71E-06
SLC2A2	54.02	-1.81	1.36E-15	0.002344939
FABP1	4382.01	-1.81	1.36E-15	0.000275653
FABP2	937.62	-1.82	1.36E-15	0.000154657
C11orf86	895.13	-1.82	1.36E-15	8.83E-07
HPGD	10579.91	-1.83	1.36E-15	2.10E-07
TUBAL3	177.38	-1.85	1.36E-15	2.12E-08
CIDEA	9.21	-1.85	1.36E-15	0.017733275
B4GALNT2	12.41	-1.86	1.36E-15	0.001563814
ADH7	318.50	-1.87	1.36E-15	0.022531514
MEP1A	4580.09	-1.88	1.36E-15	1.51E-05
GPX3	693.63	-1.89	1.36E-15	5.55E-12
OTOP3	447.72	-1.90	1.36E-15	0.000220211

SLC25A48	8.38	-1.91	1.36E-15	1.43E-07
MROH9	27.21	-1.91	1.36E-15	3.54E-06
SPATA31E1	6.86	-1.92	1.36E-15	0.000256653
SPAM1	3.52	-1.93	1.36E-15	0.001881327
RAB37	163.64	-1.93	1.36E-15	6.95E-12
SLC7A9	207.59	-1.93	1.36E-15	2.63E-06
CHP2	942.11	-1.93	1.36E-15	0.000492341
NAT8	40.12	-1.95	1.36E-15	2.74E-06
UNC5D	11.62	-1.95	1.36E-15	4.32E-08
VSTM2A	33.44	-1.96	1.36E-15	4.85E-06
SPIB	124.73	-1.96	1.36E-15	4.11E-06
ABCG2	947.10	-2.01	1.36E-15	3.75E-08
BEST4	89.12	-2.02	1.36E-15	2.45E-08
SLC51B	73.56	-2.02	1.36E-15	8.99E-09
MYH7	2.26	-2.03	1.36E-15	0.000211235
SI	13658.97	-2.03	1.36E-15	0.000130361
ALPI	150.54	-2.07	1.36E-15	1.31E-05
NR1H4	156.99	-2.08	1.36E-15	9.98E-07
SLC26A3	2816.50	-2.13	1.36E-15	7.18E-05
GKN2	2553.99	-2.14	1.36E-15	0.016462263
MAL	1505.33	-2.19	1.36E-15	0.017324313
DAB1	37.28	-2.20	1.36E-15	3.98E-07
RNASE7	123.53	-2.22	1.36E-15	0.029117983
SLC28A1	32.43	-2.22	1.36E-15	0.000849066
ADH4	1982.98	-2.25	1.36E-15	1.54E-05
SLC10A2	28.16	-2.26	1.36E-15	0.000430827
CA7	30.04	-2.27	1.36E-15	9.79E-12
CYP3A4	575.46	-2.30	1.36E-15	4.02E-05
CLCA1	745.96	-2.30	1.36E-15	0.000285807
S100A2	1920.01	-2.30	1.36E-15	0.002150957
ALDOB	3934.52	-2.31	1.36E-15	1.00E-06
SLC28A2	2562.41	-2.31	1.36E-15	4.29E-06
SLC15A1	516.55	-2.33	1.36E-15	8.83E-07
SLURP1	167.31	-2.33	1.36E-15	0.026543209
SLC2A5	136.82	-2.33	1.36E-15	9.88E-07
LINC00955	19.52	-2.33	1.36E-15	0.001370772
MTTP	507.21	-2.35	1.36E-15	0.000124018
GUCA2A	91.82	-2.37	1.36E-15	0.000101046
NKX6-2	44.66	-2.46	1.36E-15	8.68E-06
GUCA2B	61.50	-2.52	1.36E-15	1.03E-06
APOA1	272.56	-2.52	1.36E-15	3.60E-05
MUC7	16.95	-2.52	1.36E-15	0.003295514

PPY	1.76	-2.56	1.36E-15	0.017646941
MEP1B	621.90	-2.63	1.36E-15	4.32E-08
C14orf180	42.15	-2.64	1.36E-15	1.00E-09
GKN1	9386.72	-2.67	1.36E-15	0.004640015
CPA2	6.78	-2.77	1.36E-15	0.006499887
LYPD8	7.54	-2.77	1.36E-15	0.001899525
APOB	4065.05	-2.87	1.36E-15	0.000140478
GCG	77.47	-2.93	1.36E-15	0.005701052
G6PC	7.27	-2.93	1.36E-15	0.001279368
ZG16	387.60	-3.08	1.36E-15	3.03E-08
INSL5	2.43	-3.09	1.36E-15	0.000207025
CA1	2052.34	-3.11	1.36E-15	1.91E-12
GATA5	23.06	-3.11	1.36E-15	0.012234372
TMPRSS15	1848.41	-3.26	1.36E-15	1.54E-05
PGA3	207.08	-3.30	1.36E-15	2.84E-05
TMIGD1	5.14	-3.30	1.36E-15	0.002257418
SLC17A8	11.55	-3.41	1.36E-15	2.67E-05
PGA5	72.77	-3.51	1.36E-15	4.28E-05
PGA4	161.27	-3.56	1.36E-15	2.20E-05
APOC3	68.32	-3.72	1.36E-15	3.31E-05
GIF	52.07	-3.84	1.36E-15	2.12E-10
MUC22	45.18	-4.37	1.36E-15	0.015581525
ATP4B	28.62	-4.42	7.58E-16	7.34E-13
ATP4A	79.67	-5.03	1.32E-16	1.77E-13
KRT77	0.90	-6.05	1.36E-15	0.00263354